Coreferentie van atomaire en complexe objecten

Antoon Bronselaer

Promotor: prof. dr. G. De Tré
Proefschrift ingediend tot het behalen van de graad van
Doctor in de Ingenieurswetenschappen: Computerwetenschappen

Vakgroep Telecommunicatie en Informatieverwerking
Voorzitter: prof. dr. ir. H. Bruneel
Faculteit Ingenieurswetenschappen
Academiejaar 2010 - 2011

UNIVERSITEIT
GENT

ISBN 978-90-8578-392-3
NUR 993
Wettelijk depot: D/2010/10.500/68

Leden van de examencommissie

prof. dr. ir. Daniël De Zutter	(Universiteit Gent, Voorzitter)
em. prof. dr. ir. Patrick Bergmans	(Universiteit Gent, Secretaris)
prof. dr. Guy De Tré	(Universiteit Gent, Promotor)
prof. dr. ir. Maurice Van Keulen	(Universiteit Twente)
dr. Dirk Van Hyfte	(i.Know)
prof. dr. Bernard De Baets	(Universiteit Gent)
prof. dr. ir. Gert De Cooman	(Universiteit Gent)
prof. dr. ir. Herwig Bruneel	(Universiteit Gent)

Affiliaties

Onderzoeksgroep Database, Document en Content Management (DDCM)
Vakgroep Telecommunicatie en Informatieverwerking (TELIN)
Faculteit Ingenieurswetenschappen
Universiteit Gent
Sint-Pietersnieuwstraat 41
B-9000 Gent
België

Samenvatting

In hedendaagse informatiesystemen is het herkennen en verwerken van dubbele data een belangrijke uitdaging met verschillende toepassingen. In de context van databanken en *data warehouses* bijvoorbeeld, leidt dubbele data tot inconsistenties, inefficiënte opslag van data en foutieve statistieken. Bij een vergelijkende studie van concurrerende verkoopswebsites is de overlap in aangeboden producten een belangrijke indicator. Bij identificatie van personen zijn biometrische data of metadata gegeven en willen we weten welke data aan elkaar gekoppeld kunnen worden. Ook bij het beheer van multimedia zoals tekst, beeld en muziek, is het herkennen van dubbele data belangrijk.

In deze thesis beschrijven we het probleem van dubbele data in termen van objecten. Een object is een datafragment dat een entiteit uit de realiteit beschrijft. Objecten kunnen bijvoorbeeld beschrijvingen zijn van personen, auto's, huizen ... Er wordt een abstractie gemaakt van objectcreatie, d.i. het proces dat objecten aanmaakt en ze in een databank plaatst. Meer bepaald veronderstellen we dat een object het gevolg is van een meting van een entiteit. Het resultaat van deze meting wordt uitgedrukt in een objectruimte. Vervolgens argumenteren we dat meetprocessen in realiteit onderhevig zijn aan imperfecties, waardoor herhaalde metingen van dezelfde entiteit kunnen leiden tot verschillende objecten. Wanneer twee (eventueel verschillende) objecten dezelfde entiteit beschrijven, spreken we van coreferente objecten. Hoewel coreferentie van twee objecten een Boolees probleem is, hebben de meetimperfecties tot gevolg dat er onzekerheid over coreferentie van twee objecten kan ontstaan. In deze thesis geven we argumenten om deze onzekerheid te modelleren aan de hand van de possibiliteitstheorie. Meer bepaald willen we voor twee objecten een possibiliteitsverdeling over het domein van Boolese waarheidswaarden $\mathbb{B} = \{T, F\}$ construeren. Een dergelijke verdeling wordt een possibilistische waarheidswaarde genoemd en een operator die ze construeert voor twee gegeven objecten wordt een evaluator genoemd.

Het onderzoek naar evaluatoren dat in deze thesis wordt gerapporteerd, heeft volgende resultaten opgeleverd. Er is uitgegaan van een coreferentierelatie die moet worden geconstrueerd. Aangezien coreferentie van objecten voortkomt uit gelijkheid van de beschreven entiteiten, is de coreferentierelatie een equivalentierelatie. Er is bijgevolg bestudeerd in welke mate de eigenschappen van een equivalentierelatie restricties opleggen aan evaluatoren. Er is aangetoond

hoe een coreferentierelatie kan worden opgebouwd op basis van een evaluator. Hierbij wordt uitgegaan van beslissingsmodellen. Tweewaardige beslissingsmodellen beelden een possibilistische waarheidswaarde af op een tweewaardige ruimte ($\mathbb{B}$) en nemen een Boolese beslissing op basis van de gegeven onzekerheid. Driewaardige beslissingsmodellen laten toe om in situaties van hoge onzekerheid geen beslissing te nemen. We hebben aangetoond hoe op basis van beslissingsmodellen een consistente coreferentierelatie kan worden opgebouwd. Er is mede aangetoond hoe bepaalde oorzaken van onzekerheid onafhankelijk zijn van het universum waarin objecten worden voorgesteld. Een eerste voorbeeld hiervan is het probleem van moeilijk-meetbaarheid. Om dit probleem op te lossen definiëren we een evaluator voor possibilistische variabelen. Een tweede voorbeeld hiervan is het probleem van semantische evaluatoren. Semantische evaluatoren gebruiken een binaire relatie over de objectruimte om onzekerheid over de coreferentie van twee objecten af te leiden. Daar tegenover staan syntactische evaluatoren die onzekerheid afleiden op basis van een syntactische vergelijking van twee objecten. In onze studie van syntactische evaluatoren definiëren we eerst evaluatoren voor (multi)verzamelingen. Dergelijke evaluatoren houden expliciet rekening met het feit dat coreferentie van (multi)verzamelingen niet kan worden afgeleid door verificatie van de gelijkheid van elementen. Om die reden wordt een evaluator op het niveau van elementen beschouwd. Op basis van deze evaluator wordt een leximax-optimale één-op-één afbeelding tussen (multi)verzamelingen geconstrueerd. We introduceren een algoritme dat een dergelijke afbeelding construeert en we analyseren de complexiteit van dit algoritme. Bovendien worden de eigenschappen van onze aanpak onderzocht. We passen de resultaten toe in het geval van karakterstrings. Hierbij wordt een splitsingsfunctie gebruikt die een karakterstring omzet naar een multiverzameling van deelstrings. Naast enkele theoretische voordelen tonen experimenten aan dat onze aanpak een verbetering biedt ten opzichte van bestaande methoden.

De syntactische evaluatoren voor (multi)verzamelingen en karakterstrings die worden gedefinieerd, steunen op een nieuwe combinatiefunctie voor possibilistische waarheidswaarden, namelijk de Sugeno-integraal voor possibilistische waarheidswaarden. Deze functie is gebaseerd op twee vertrouwensmaten die samen conditionele necessiteit modelleren. We tonen aan hoe deze nieuwe functie zich verhoudt tot bestaande combinatiefuncties. Zo argumenteren we dat conditionele necessiteit overeenkomt met een soort van afhankelijkheid, die orthogonaal staat op het concept van t-onafhankelijkheid. We onderzoeken de eigenschappen van onze combinatiefunctie en we bekijken enkele bijzondere gevallen. Naast het gebruik van de nieuwe combinatiefunctie bij de constructie van evaluatoren voor (multi)verzamelingen en karakterstrings, gebruiken we de combinatiefunctie ook om evaluatoren voor complexe objecten te definiëren. Complexe objecten beschrijven een entiteit aan de hand van beschrijvingen van eigenschappen van de entiteit. Uit experimentele resultaten blijkt dat onze evaluatoren voor complexe objecten beter presteren dan bestaande methoden.

De vermelde evaluatoren voor karakterstrings zijn bruikbaar in een context

van complexe objecten. Daarnaast hebben we ook onderzocht hoe karakterstrings behandeld moeten worden wanneer ze een losse, tekstuele beschrijving van een entiteit voorstellen. Dit leidt naar een nieuwe aanpak voor tekstclustering, waarbij we gebruik maken van de vernieuwende i.Know technologie om een nieuw tekstmodel op te stellen. Dit nieuwe model is gebaseerd op multirelaties en wordt daarom het relationele tekstmodel genoemd. We definiëren operatoren voor dit model die ons toelaten om snel en efficiënt teksten te groeperen in clusters. Ook laten we het verband zien met het raamwerk van evaluatoren. We geven een methode voor het bepalen van het aantal entiteiten dat wordt gerefereerd. We tonen aan hoe onze methode het resultaat is van een afweging tussen *zuiverheid* en *compleetheid*. Experimentele resultaten wijzen op een sterke verbetering ten opzichte van bestaande methoden.

Ten slotte is in dit werk ook onderzocht hoe coreferente objecten verwerkt kunnen worden. Dit is bijvoorbeeld nodig in de context van databanken en *data warehouses*, waarbij een groep van coreferente objecten moet worden samengevoegd tot één object. Hierbij is het belangrijk dat dit ene object een zo goed mogelijke beschrijving is van de entiteit in kwestie. Daarom wordt een beginnende studie gemaakt van samenvoegingsfuncties die een multiverzameling van coreferente objecten samenvoegen tot één object. Er worden samenvoegingsfuncties voor atomaire objecten voorgesteld op basis van een evaluator en we laten zien hoe dergelijke functies kunnen worden samengesteld tot samenvoegingsfuncties voor complexe objecten.

Summary

In modern information systems, the recognition and processing of duplicate data is an important challenge with several applications. In the context of databases and data warehouses for example, duplicate data lead tot inconsistencies, inefficient data storage and erroneous statistics. When making a comparative study of competing sales websites, the overlap of offered products is an important indicator. The problem of people identification deals with comparison of biometrical data or metadata. Also, the management of multimedia such as texts, images and music gains benefit from duplicate data recognition.

In this thesis we describe the problem of duplicate data in terms of objects. An object is a fragment of data that describes an entity from the real world. For example, objects can describe persons, cars, real estate ... An abstraction of object creation is made. Object creation is the process that creates objects and puts them in a database. It is assumed that an object is the result of a measurement of an entity. This result is expressed in an object space. We claim that such measurements can suffer from imperfections, which is why different measurements of the same entity can lead to different objects. Objects that describe the same entity are called coreferent objects. Although coreference of objects is a Boolean matter, imperfections of the measurement imply that uncertainty about coreference can exist. In this work, such uncertainty is treated by means of possibility theory. More precise, for two objects, we construct a possibility distribution over the domain of Boolean truth values $\mathbb{B} = \{T, F\}$. Such a distribution is called a possibilistic truth value and an operator that constructs it, is called an evaluator.

The study of evaluators reported in this thesis leads to the following results. It is assumed that a coreference relation must be constructed. Considering the fact that coreference of objects is a matter of equality of the described entities, a coreference relation must be an equivalence relation. It is therefor verified to what extent an evaluator is restricted by the properties of an equivalence relation. It is shown how a coreference relation can be constructed, based on an evaluator. For that purpose, decision models are defined. A binary decision model maps a possibilistic truth value to a binary space ($\mathbb{B}$). This means that a Boolean decision is taken based on the given uncertainty. A ternary decision model allows that, in the case of high uncertainty, a decision is avoided. We show how decision models can be used to construct a consistent

coreference relation. It is then shown that some causes of uncertainty are independent of the universe in which objects are represented. A first example hereof, is the problem of difficult measurement. In order to solve this problem, we define an evaluator for possibilistic variables. A second example is the problem of semantical evaluation. Semantical evaluators use a binary relation over the object space to infer uncertainty about coreference for two objects. Opposed to that, syntactical evaluators infer uncertainty based on a syntactical comparison of two objects. In our study of syntactical evaluators, we first define evaluators for (multi)sets. These evaluators take into account that coreference of (multi)sets cannot be determined by verification of the equality of elements. Therefor, an evaluator on the level of elements is taken into account. This evaluator is used to construct a leximax-optimal one-to-one mapping between (multi)sets. We introduce an algorithm that is able to construct such a mapping and an analysis of the complexity of this algorithm is given. The properties of our approach are investigated. We apply the results about comparison of sets and multisets to the case of strings. A split function is used that transforms a string into a multiset of substrings. Next to some theoretical advantages of our approach, experiments show that our approach offers an improvement with respect to existing methods.

The syntactical evaluators for (multi)sets and strings use a novel combination function for possibilistic truth values called the Sugeno-integral for possibilistic truth values. This function is based on two confidence measures that model conditional necessity. It is shown how this novel combination function behaves with respect to existing combination functions. For example, we show that conditional necessity comes from a type of dependencies that is orthogonal to the concept of t-independence. We investigate the properties of our combination function and some special cases are studied. Next to the use of the combination function in the construction of syntactical evaluators for strings, we also use it for the construction of evaluators for complex objects. A complex object describes an entity by means of descriptions of properties of the entity. Experimental results show that our evaluators for complex objects perform better than existing methods.

The mentioned evaluators for strings are useful in the context of complex objects. Next to this context, strings can also be used to provide a textual description of an entity. In order to deal with this case, a new approach for text clustering is proposed. Our approach uses the innovative technology by i.Know to create a novel model for text that is based on multirelations. We call this model the relational model for text. We define operators that allow us to quickly group texts into clusters. We show the connection between our approach for text clustering and the framework of evaluators. A method for the estimation of the number of described entities is given. We show that our method is the result of a balance between precision and recall. Experimental results indicate a strong improvement compared to existing methods.

Finally, this thesis also contributes to the processing of coreferent objects. In the context of databases and data warehouses, it is required that coreferent

objects are merged into one object. This object must be an optimal description of the entity that is dealt with. For that reason, a preliminary framework of merge functions is introduced. Hereby, a merge function is a function that maps a multiset of objects to a single object. Merge functions for atomic objects are proposed based on an evaluator. We show how these functions can be composed into merge functions for complex objects.

Dankwoord

"Begegnet uns jemand, der uns Dank schuldig ist, gleich fällt es uns ein. Wie oft können wir jemand begegnen, dem wir Dank schuldig sind, ohne daran zu denken."

–Johann Wolfgang von Goethe, 1749-1832

De hierop volgende driehonderd bladzijden bevatten een samenvatting van vier jaar onderzoek, maar het dubbele volstaat niet om mijn diepste dank te betuigen aan hij of zij die dat verdient. Een notie van realiteit dwingt mij echter tot een sterk ingekorte versie.

Ik wens in de eerste plaats mijn promotor Guy De Tré te bedanken voor de kans die ik van hem heb gekregen om onderzoek te verrichten. Ik heb een grote appreciatie voor de technische discussies, voor de vele niet-technische discussies en voor het grote geduld waarmee hij mij heeft opgeleid tot de onderzoeker die ik vandaag ben. Nauw aansluitend daarbij wil ik ook mijn collega's bij TELIN en in het bijzonder de mensen van DDCM (Axel, Bert, Christophe, Daan, Joachim, Jörg en Tom) bedanken voor de leuke tijd. Ook wil ik de leden van de jury bedanken voor het grondig evalueren van dit werk en prof. De Cooman en prof. De Baets voor hun interessante suggesties vooraf. In het kader van de praktische toepassing van dit werk ben ik dank verschuldigd aan Saskia Debergh, dr. Dirk Van Hyfte en alle mensen van i.Know en tevens aan Joan De Winne en alle mensen van DVI.

Al sinds lang voor ik aan dit werk begon, kan ik rekenen op de steun van heel wat vrienden en familie. Ik dank in het bijzonder mijn goede vriend Maarten en met een zekere tristesse ook mijn goede vriendin Lena voor hun jarenlange vriendschap. Ik hoop, Lena, dat je de rust hebt gevonden die je zo hard nodig had. Ik wil graag Freija, Dorien, Wim, Koen en Heidi bedanken voor de leuke babbels.

Mijn muzikale kornuiten van Mr. Panter zijnde Koen, Bart, Fred, Wob, Pieter-Jan en vroeger ook Chris dank ik voor de leukste muziek die ik ken, de E.P. Silence in Stereo en vooral het vele plezier onderweg.

Ik wil graag ook mijn familie en schoonfamilie bedanken voor wie ze zijn, niet in het minst mijn mama en papa, zonder wie ik niemand zou zijn. Ik kijk met plezier terug op mijn jeugd en een warme thuis. Ik draag ook mijn grootouders, mijn grote broer Joost en zijn vriendin Edith, mijn lieve zus Veerle en

haar vriend Kris, mijn meter Magda en peter Fernand, mijn neven en nichten, mijn schoonouders en mijn “roomie” Joke een warm hart toe.

Als laatste rest nog de essentie. Mijn diepste liefde en genegenheid gaat uit naar mijn hartsvriendin Sarah. Al ruim vijf jaar ben je mijn steun en toeverlaat, mijn lachen en huilen, mijn andere “ik”. Je bent meer dan ik kan wensen.

Inhoudsopgave

Publicaties

Artikels in tijdschriften opgenomen in de Science Citation index, Social Science Citation Index of Arts en Humanities Citation Index (A1)

1. Bronselaer Antoon and Guy De Tré. Properties of possibilistic string comparison. *IEEE Transactions on Fuzzy Systems*, 18 (2010): 312-325. 2009

2. Bronselaer, Antoon, Axel Hallez and Guy De Tré. A possibilistic view on set and multiset comparison. *Control and Cybernetics*, 38 (2009): 341-366.

3. Hallez Axel, Guy De Tré and Antoon Bronselaer. Performance optimization of object comparison. *International Journal of Intelligent Systems*, 24 (2009): 1057-1076.

4. Hallez Axel, Antoon Bronselaer and Guy De Tré. Comparison of sets and multisets. *International Journal of Uncertainty Fuzziness and Knowledge-based Systems*, 17 (2009): 153-172.

5. Bronselaer Antoon, Axel Hallez and Guy De Tré. Extensions of fuzzy measures and Sugeno integral for possibilistic truth values. *International Journal of Intelligent Systems*, 24 (2009): 97-117.

6. Bronselaer Antoon and Guy De Tré. A possibilistic approach to string comparison. *IEEE Transactions on Fuzzy Systems*, 17 (2009): 208-223.

Artikels in proceedings van wetenschappelijke conferenties, opgenomen in de Science Citation index, Social Science Citation Index of Arts en Humanities Citation Index (P1)

1. Bronselaer Antoon and Guy De Tré. Aspects of object merging. In *Proceedings of the NAFIPS Conference*, 27-32. Toronto, Canada, 2010.

2. De Tré Guy and Antoon Bronselaer. Consistently handling geographical user data: Merging of coreferent POIs. In *Proceedings of the NAFIPS Conference*, 117-122. Toronto, Canada, 2010.

3. De Tré Guy, Antoon Bronselaer, Tom Matthé, Nico Van de Weghe and Philippe De Maeyer. Consistently Handling Geographical User Data: Context-Dependent Detection of Co-located POIs. In *Communications in Computer and Information Science*, edited by Eyke Hüllermeier, Rudolf Kruse and Frank Hoffmann, 85-94. Vol. 81. 13th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, IPMU 2010 81. Dortmund, Germany: Springer-Verlag, 2010.

4. Bronselaer Antoon, and Guy De Tré. Impact of [0,1]-valued weights and weighted aggregation operators for possibilistic truth values. In 2008 Annual meeting of the north american fuzzy information processing society, VOLS 1 and 2, 162-167. IEEE, 2008.

Artikels in proceedings van wetenschappelijke conferenties, niet opgenomen in voorgaande secties (C1)

1. Bronselaer Antoon, Saskia Debergh, Dirk Van Hyfte and Guy De Tré. Estimation of topic cardinality in document collections. In *Towards Natural Language Based Data/Text Mining and Summarization via Soft Approaches*, edited by Janusz Kacprzyk and Slawomir Zadrozny, 31-39. SIAM, 2010.

2. De Tré Guy, Slawomir Zadrozny, Tom Matthé, Janusz Kacprzyk and Antoon Bronselaer. Dealing with positive and negative query criteria in fuzzy database querying : bipolar satisfaction degrees. In *Lecture notes in Computer Science*, edited by Andreasen Troels, 593-604. Vol. 5822. Springer Verlag Berlin, 2009.

3. Bronselaer, Antoon, and Guy De Tré. Semantical evaluators. In *Proceedings of the 2009 IFSA World Congress / EUSFLAT Conference*, 2009.

4. Bronselaer Antoon, Axel Hallez, and Guy De Tré. Evaluation in the possibilistic framework for object matching. In *IPMU 2008*, 2008.

5. Bronselaer Antoon, Joan De Winne, and Guy De Tré. Flexible matching of Ear Biometrics. In *Proceedings of the First International VLDB Workshop on Management of Uncertain Data (MUD)*, 5-17., 2007.

6. Bronselaer Antoon, Guy De Tré, and Axel Hallez. Dynamic preference modeling in flexible object matching. *In Proceedings EuroFuse Workshop: New Trends in Preference Modelling*, 191-195., 2007.

Lijst van figuren

Voorbeschouwing

Inleiding

Sinds de ontwikkeling van de transistor in 1947 is de mondiale samenleving in een razendsnel tempo uitgegroeid tot een volwaardige informatiemaatschappij. Het contrast tussen de eerst gekende digitale rekenmachine (Harvard Mark I) met een massa van ongeveer vijf ton en het hedendaags onderzoek naar nanotechnologie is haast niet te vatten voor de menselijke geest. De verschroeiende technologische vooruitgang wordt vandaag de dag nog steeds bevattelijk omschreven door de Wet van Moore, die stelt dat het aantal transistoren op een computerchip elk jaar verdubbelt. Hoewel de exponenten van de informatiemaatschappij in haar huidige vorm alomgekende applicaties zoals het *Wereld-Wijde Web* (WWW), sociale netwerksites en *e-mail* zijn, behelst de term in essentie veel meer dan dat. Een informatiemaatschappij wordt volgens Van Dale gedefinieerd als "een samenleving waarin de informatievoorziening en het beheersen van de informatiestromen essentieel zijn". Deze definitie impliceert dat informatie zowel een sociaal als economisch draagvlak vormt. De geschiedenis leert dat de uitvinding van de computer als automatische verwerkingseenheid van informatie voor een groot deel is versneld door militaire doelstellingen in de Tweede Wereldoorlog. De Harvard Mark I werd immers gebruikt om het traject van raketten te berekenen en is een voorbeeld van de voorsprong die informatie kan bieden. De vaststelling dat informatie equivalent is aan voorsprong speelt zonder twijfel een belangrijke rol in de historische ontwikkelingen op gebied van informatieopslag en informatieverwerking. Dit principe is vandaag meer dan ooit van toepassing in een economie waar concurrentie meer en meer gedreven wordt door optimaal beheer en uitbaten van kennis en informatie. Hoewel dit werk geen sociologische analyse is van de informatiemaatschappij, is het globale belang van informatie binnen onze maatschappij duidelijk een fundamenteel argument voor dit werk.

De groeiende mogelijkheden voor opslag van data hebben in de tweede helft van de 20^{ste} eeuw de noodzaak aan structuur in de opslag van data sterk doen toenemen. Het onderzoek naar deze problematiek kent in 1970 een doorbraak met de ontwikkeling van het relationele datamodel door Codd [1]. Het belang van dit model is tweevoudig. Enerzijds is het relationele model nog steeds hét toonaangevende model voor moderne *databanken*. Anderzijds is dit model

tekenend voor het inzicht dat wiskundige en abstracte modellen een zeer belangrijke rol kunnen spelen als theoretisch fundament voor de ontwikkeling van software. Een voorbeeld hiervan is de λ-calculus [2], die een belangrijke theoretische onderbouw geeft voor programmeertalen. Een veelheid aan hedendaagse onderzoeksgebieden zijn ontstaan uit het besef dat praktische applicaties hun voordeel halen uit wiskundige modellering.

Daar waar het relationele datamodel een standaard vormt voor de opslag van data, betekende de ontwikkeling van het WWW een vooruitgang in het verspreiden van data. De combinatie van databank en netwerktechnologie heeft gezorgd voor het toegankelijk maken van een alsmaar groeiende hoeveelheid data voor particuliere en professionele doeleinden. De problematiek die gekoppeld is aan zulke oncontroleerbaar grote hoeveelheden data is sinds eind vorige eeuw een belangrijke onderzoeksuitdaging die heeft geleid tot het ontstaan van *datamining*, met sterke toepassingen in het *Business Intelligence*(BI) proces. In dit proces wordt informatie over economische activiteiten gegenereerd uit de bedrijfsdata met als doel de optimalisatie van de bedrijfsstrategie. Het steeds belangrijker worden van dit proces is op zich een kenmerk van de informatiemaatschappij.

Binnen de problematiek die gepaard gaat met het beheer van grote hoeveelheden data, bestaan een aantal problemen waarbij gegevens onderling vergeleken moeten worden, of waarbij het vergelijken van gegevens minstens aanleiding geeft tot een mogelijke oplossing van het probleem. De voorbeelden van dergelijke problemen zijn uiteenlopend van aard:

- In identificatietaken wordt een (meestal biometrische) stempel van een referentiepersoon vergeleken met een groep van andere personen, met als doel de identiteit van de referentiepersoon te achterhalen (vingerafdrukken, retina-scans, paswoorden ...).

- Bij beslissingsproblemen worden een aantal oplossingen vooropgesteld waarvan dient uitgemaakt te worden welke de beste is. Een eerste aanpak hiervoor is de onderlinge vergelijking van de verschillende oplossingen, leidende tot een preferentierelatie over de oplossingsruimte. Een tweede aanpak zal een ideale (niet noodzakelijk bestaande) oplossing bepalen en toetsen welke bestaande oplossing het beste aansluit bij de ideale oplossing. Dit probleem heeft relevante toepassingen bij aankopen allerhande (televisietoestel, huis ...).

- Bij het gebruik van zoekmachines zoals Google op het WWW worden documenten vergeleken met een verzameling van trefwoorden.

- Het tekstclusteringsprobleem handelt over het opsplitsen van een verzameling van teksten in groepen, zodat teksten in eenzelfde groep over eenzelfde onderwerp handelen. Hierbij moeten teksten onderling vergeleken worden op basis van het onderwerp dat ze behandelen.

- Gegeven een klantenbestand van een bedrijf, zoek de klanten die dure producten gekocht hebben tijdens de voorbije drie maanden. Een der-

gelijk probleem wordt typisch opgelost door een query te formuleren die zoekcriteria vergelijkt met gegevens uit een databank.

In het formuleren van een oplossing voor beslissingsproblemen en voorspellingsproblemen, leert de geschiedenis dat basismodellen voor deze problemen meestal vertrekken van technieken waarbij onderlinge vergelijking van data fundamenteel is. In de context van clustering zijn hiërarchische clustertechnieken de absolute basistechnieken en in de context van classificitie is het *k-nearest-neighbour* algoritme een welgekende standaardmethode. Deze vaststelling kan worden verklaard door te stellen dat een vergelijkende één-aan-één studie van oplossingen vaak een zeer intuïtieve methode is om problemen op te lossen. Hoewel de meest intuïtieve oplossing geen garantie biedt op de beste oplossing, mag het intuïtieve karakter van oplossingsmethoden niet onderschat worden. Het is een ongeschreven wetmatigheid in de wereld van *datamining* dat een techniek pas aanslaat wanneer mensen met een niet technische achtergrond deze techniek intuïtief aanvaarden, hetgeen bijvoorbeeld de populariteit van beslissingsbomen verklaart.

Probleemstelling

Het probleem dat in deze thesis wordt bestudeerd, kadert binnen de voorgaande uiteenzetting van enerzijds massale hoeveelheden data en anderzijds het vergelijken van gegevens. Meer specifiek worden in deze thesis oplossingen aangereikt voor het coreferentieprobleem. Dit houdt in dat we zoeken naar koppels van gegevensstructuren die, onafhankelijk van elkaar, eenzelfde entiteit uit de reële wereld beschrijven. We zullen hierbij een onderscheid maken tussen het voorkomen van dit probleem binnen enerzijds goed gestructureerde databronnen zoals databanken en anderzijds databronnen waarbij enkel een vrije en informele beschrijving voorhanden is. Bij de eerste beschouwing van het probleem wordt uitgegaan van een vooraf gedefinieerde beschrijving van de data die wordt ingevuld volgens een specifiek vraag-antwoord principe. Bijvoorbeeld, in een databanktabel die persoonsbeschrijvingen bijhoudt, bevat het veld 'voornaam' het antwoord op de vraag: "Wat is de voornaam van persoon x?". Er wordt onderzocht hoe gestructureerde gegevens die eenzelfde entiteit voorstellen, aan elkaar gelinkt kunnen worden. De oplossing voor dit probleem kan worden toegepast in onder andere identificatieproblemen, databankfilters voor het opkuisen van databanken en *Extract Transform Load* processen (ETL) waarbij verschillende operationele databanken samengebracht worden in één *data warehouse*. De tweede beschouwing van het probleem veronderstelt een vrije, tekstuele beschrijving van entiteiten. Het invullen van data gebeurt hier volgens een ruimer principe, met vragen in de vorm van "Hoe zou u persoon x beschrijven?". Er wordt een methode geconstrueerd die uitgaat van een reeks databronnen die, al dan niet onafhankelijk, antwoorden genereren op dergelijke vragen. Het coreferentieprobleem manifesteert zich dan als het zoeken naar antwoorden waarvoor de vraag dezelfde is. Vanuit deze invalshoek zal een nieuw model worden opgebouwd dat gezien kan worden als een mechanisme

voor clustering van tekstuele beschrijvingen. Beide problemen worden eerst vanuit een theoretisch standpunt bekeken, hetgeen leidt tot een theoretisch onderbouwde oplossing. Voorgestelde oplossingen worden telkens getoetst op hun effectiviteit door vergelijking met bestaande aanpakken van de problemen in kwestie.

Doelstelling

De concrete doelstelling van deze thesis is het onderzoeken en construeren van een raamwerk waarbinnen het coreferentieprobleem in zijn verschillende instanties kan worden aangepakt. Ten eerste moet deze thesis een bijdrage leveren tot het onderzoeken en toepassen van possibiliteitstheorie op het probleem van gestructureerde data. Ten tweede wordt onderzocht hoe het probleem van tekstuele beschrijvingen kan worden opgelost door te vertrekken van een fundamenteel nieuwe voorstellingswijze van teksten. We willen op dit vlak aantonen dat, mits de correcte voorstellingswijze, eenvoudige methoden kunnen leiden tot bijzonder goede resultaten. Ten derde willen we steeds komen tot concrete oplossingen voor de gestelde problemen door uitwerking van operatoren voor vergelijking van objecten. We zullen deze oplossingen vergelijken met methoden uit de literatuur. Om deze vergelijking mogelijk te maken zal een prototype software worden ontwikkeld die verder kan worden gebruikt als de basis voor toepasbare en commerciële softwarepakketten.

Overzicht

Deze thesis is als volgt ingedeeld. In het inleidende Hoofdstuk 1 worden basisnotaties ingevoerd en wordt een overzicht gegeven van de belangrijkste literatuur die aan de basis van deze thesis ligt. Hierbij wordt voornamelijk aandacht besteed aan onzekerheidstheorie. In Hoofdstuk 2 formaliseren we het coreferentieprobleem en wordt een analyse gemaakt van deze probleemstelling. Enerzijds wordt besproken wat de oorzaak van coreferentie is en anderzijds wordt bekeken wat de preventieve maatregelen kunnen zijn om het coreferentieprobleem te vermijden. Voorts behandelt Hoofdstuk 2 de definitie, eigenschappen en voorbeelden van een fundamentele operator binnen deze thesis, namelijk een evaluator. In Hoofdstuk 3 wordt onderzocht hoe kennis uit verschillende bronnen kan worden gecombineerd. Hierbij wordt uitgegaan van een aanpak met vertrouwensmaten. In Hoofdstuk 4 worden de resultaten uit Hoofdstuk 3 gebruikt om te komen tot een possibilistische aanpak voor het vergelijken van collecties. De eigenschappen van deze methode worden grondig onderzocht. In Hoofdstuk 5 worden beslissingsmodellen gedefinieerd als een middel om de resultaten van evaluatoren om te zetten tot een concrete oplossing voor het coreferentieprobleem. Hierbij wordt aangetoond hoe inconsistenties in de resultaten kunnen worden behandeld en hoe aan randvoorwaarden voldaan kan worden. Hierbij worden de resultaten uit Hoofdstuk 4 gebruikt. De resultaten

uit Hoofdstukken 3, 4 en 5 worden samengebracht in Hoofdstuk 6 om te komen tot een possibilistische methode voor de vergelijking van karakterstrings. Hierbij wordt eerst een formalisatie van karakterstrings voorgesteld. Daarna worden zowel één-niveau als twee-niveau methoden geïntroduceerd vanuit een possibilistisch standpunt. De twee-niveau methode steunt op een geparameteriseerde functie en er wordt bekeken hoe de parameters bepaald kunnen worden, zowel met als zonder trainingsdata. Ten slotte biedt Hoofdstuk 6 een uitgebreide experimentele evaluatie van de voorgestelde technieken. In Hoofdstuk 7 wordt een possibilistische methode gegeven voor het zoeken naar coreferente complexe objecten. Hierbij wordt enerzijds uitgegaan van de evaluatoren uit eerdere hoofdstukken en anderzijds van de technieken voor combinatie van kennis uit Hoofdstuk 3. Er wordt aandacht besteed aan het bepalen van de vertrouwensmaten zonder het gebruik van trainingsdata. Opnieuw wordt een experimentele evaluatie gemaakt van de voorgestelde technieken. Hoofdstuk 8 richt zich op het coreferentieprobleem in de context van tekstuele data. Er wordt uitgegaan van een nieuwe voorstellingsmethode voor tekst, die steunt op gepatenteerde technologie van het bedrijf i.Know. Op basis van het nieuwe voorstellingsmodel wordt eerst onderzocht hoe het aantal onderwerpen kan worden geschat, gegeven een collectie van teksten. Daarna wordt een techniek voorgesteld voor het clusteren van tekstuele documenten. Ook hier wordt een experimentele studie verricht om het nut en de relevantie van de voorgestelde technieken te benadrukken. In Hoofdstuk 9 wordt onderzocht hoe coreferente data verder verwerkt kunnen worden. Hiervoor wordt een beginnende studie gemaakt van de eigenschappen van samenvoegingsfuncties. Vervolgens worden een aantal praktische functies voorgesteld en worden de eigenschappen van deze functies geëvalueerd.

Over de taal en de zetting

De taal van dit proefschrift is het Nederlands, hetgeen een bewuste keuze is geweest. Gelet op het belang van het Engels in een wetenschappelijke context en meer nog in de context van informatica is dit zeker geen evidente keuze. De keuze voor het Nederlands is voortgekomen uit de noodzaak om subtiele verschillen en accenten op een zo goed mogelijke wijze uit te drukken. Dit is in een vreemde taal altijd een stuk moeilijker dan in de moedertaal. In de wereld van de informatica bestaan er heel wat concepten die een standaard Engelse benaming hebben. Eerder dan deze Engelse benamingen over te nemen is hier geprobeerd de Nederlandse taal zo juist mogelijk te gebruiken en zijn vertalingen voorzien. Om verwarring te vermijden worden deze termen steeds *schuin gedrukt* en staan (standaard) Engelse benamingen vermeld in Bijlage B. Wanneer geen zinvolle vertaling gevonden is, staat de Engelse benaming (*schuin gedrukt*) vermeld in de tekst. Deze regel is niet gehandhaafd bij het neerschrijven van pseudocodefragmenten, dit uit overweging dat pseudocode veelal verwijst naar bestaande sleutelwoorden uit programmeertalen. Hoewel dergelijke sleutelwoorden altijd Engelstalige termen zijn, overstijgen ze de Engelse

taal en worden ze eerder aanvaard als een universele notatie. Voor de duidelijkheid zullen deze sleutelwoorden steeds in **vet gedrukt** staan. Wanneer we vanuit een bepaald punt in de tekst willen verwijzen naar een ander punt in de tekst, dan doen we dit telkens door de vermelding van een sleutelwoord gevolgd door een nummer. Zo kunnen we bijvoorbeeld verwijzen naar 'Hoofdstuk 3' of 'Sectie 3.1' (de eerste sectie in Hoofdstuk 3). Eenzelfde principe wordt gebruikt voor figuren en tabellen. Wanneer we verwijzen naar een wiskundige formule, dan doen we dit door een nummer tussen ronde haken. Zo verwijst (1.30) naar formule 30 uit Hoofdstuk 1. Bij het verwijzen naar formules kunnen eventueel bijkomende sleutelwoorden als 'Vergelijking' voorgaan aan het nummer. Referenties naar werken uit de literatuur gebeuren door het vermelden van een nummer tussen vierkante haken, bijvoorbeeld [20]. Dit nummer verwijst naar een bijgevoegde lijst van werken uit de wetenschappelijke literatuur. Wat betreft wiskundige notatie worden vectoren en matrices in **vet gedrukt**.

Hoofdstuk 1

Vaagverzamelingen en onzekerheidsmodellen

1.1 Inleiding

In dit inleidende hoofdstuk worden een aantal wiskundige raamwerken aangehaald die een basis vormen voor de oplossingen aangereikt in deze thesis. Eerst worden twee bijzondere uitbreidingen van Cantoriaanse verzamelingen toegelicht, namelijk vaagverzamelingen en multiverzamelingen. Hierbij worden enkele belangrijke definities en basisoperatoren geïntroduceerd. Vervolgens wordt een overzicht gegeven van raamwerken voor het modelleren van onzekerheid en in het bijzonder wordt de possibiliteitstheorie besproken. Het raamwerk van possibilistische waarheidswaarden dat gebaseerd is op deze theorie, zal een belangrijke rol spelen in het vervolg van deze thesis.

1.2 Uitbreidingen van verzamelingen

De klassieke verzamenlingenleer is ontstaan vanuit filosofische discussies over het begrip oneindigheid. De intuïtieve grondbeginselen van de verzamelingenleer zijn ontwikkeld door Cantor [3] en later geformaliseerd door Fraenkel [4] en Zermelo [5] tot wat men de 'axiomatische verzamelingenleer' noemt. De formalisering van de theorie speelt een belangrijke rol in haar aansluiting bij de moderne wiskunde. Dit is voornamelijk te wijten aan de paradoxen die zich stellen in de formulering van Cantor. Een bekend voorbeeld hiervan is de paradox van de Barbier van Sevilla die beweert alle mannen te scheren die zichzelf nooit scheren. De barbier kan echter geen antwoord formuleren op de vraag of hij zichzelf scheert.

De manier waarop verzamelingen precies worden gedefinieerd is in het kader van dit werk op zich niet zeer belangrijk. Het volstaat daarom te steunen op een informele omschrijving uit de axiomatische verzamelingenleer. Deze

omschrijving zegt dat, gegeven een universum X en een eigenschap A, dan bestaat de verzameling V van alle elementen die de eigenschap A bezitten. Deze omschrijving kan men wiskundig modelleren door de definitie van een lidmaatschapsfunctie in te voeren.

Definitie 1.1 (Lidmaatschapsfunctie)
Gegeven een universum X en een eigenschap A, dan is de lidmaatschapsfunctie voor de verzameling V gedefinieerd als:

$$\mu_V : X \to \{0,1\} : \mu_V(x) = \begin{cases} 1 & \textbf{als} \quad x \text{ voldoet aan } A \\ 0 & \textbf{als} \quad x \text{ voldoet niet aan } A. \end{cases} \tag{1.1}$$

In 1965 wordt door Zadeh een theoretische veralgemening van de verzamelingenleer voorgesteld, namelijk de vaagverzamelingenleer [6]. Zijn terechte motivatie voor dit werk is dat mensen dagelijks worden geconfronteerd met communicatie in een natuurlijke taal en dat deze taal intrinsiek onzekerheid bevat. Hij argumenteert verder dat de klassieke verzamelingenleer niet in staat is een onderbouw te geven aan deze onzekerheid. Zo is het volgens Zadeh onmogelijk de verzameling V van mooie vrouwen te definiëren in de klassieke zin van het woord 'verzameling', aangezien de eigenschap 'mooi' niet eenduidig te definiëren valt. Om deze redenen vervaagt Zadeh het concept van verzamelingen door de lidmaatschapsfunctie te veralgemenen tot $\mu_{\widetilde{V}} : X \to [0,1]$. Het symbool $\widetilde{\ }$ wordt gebruikt om het onderscheid met klassieke verzamelingen zichtbaar te houden. Een element uit het universum kan nu gradueel tot een vaagverzameling $\widetilde{V}$ behoren, waarbij de graad van lidmaatschap voor een element x wordt bepaald door $\mu_{\widetilde{V}}(x)$. Het kan worden ingezien dat een dergelijke lidmaatschapsfunctie een kwantitatieve uitdrukking geeft aan voorkeuren tussen elementen uit X. Dit is precies de reden waarom vaagverzamelingen slagen waar verzamelingen falen. Waar bij klassieke verzamelingen de kenmerkende eigenschap A wel of niet aanwezig is, wordt er bij vaagverzamelingen de vraag gesteld in welke mate de eigenschap A aanwezig is. Op die manier wordt er eigenlijk ondersteld dat er een voorkeursrelatie over X bestaat, zodat elementen in X geordend kunnen worden. In het kader van het voorbeeld van Zadeh is het inderdaad intuïtief aanvaardbaar dat iedere persoon een voorkeur kan uitdrukken over welke van twee vrouwen hij/zij mooier vindt. Een vaagverzameling wordt als volgt gedefinieerd.

Definitie 1.2 (Vaagverzameling)
Gegeven een universum X. Een vaagverzameling $\widetilde{V}$ in het universum X wordt gekarakteriseerd door de lidmaatschapsfunctie:

$$\mu_{\widetilde{V}} : X \to [0,1]. \tag{1.2}$$

Een vaagverzameling $\widetilde{V}$ is genormaliseerd als:

$$\max_{x \in X} \mu_{\widetilde{V}}(x) = 1. \tag{1.3}$$

In wat volgt zal steeds worden verondersteld dat een vaagverzameling genormaliseerd is. De verzameling van alle genormaliseerde vaagverzamelingen over een universum X wordt hier genoteerd als $\mathcal{F}(X)$. Zadeh geeft naast een filosofische staving voor zijn conceptueel idee tevens een basisraamwerk voor bewerkingen van vaagverzamelingen. Een vaagverzameling wordt leeg genoemd als er geldt dat:

$$\forall x \in X : \mu_{\widetilde{V}}(x) = 0. \tag{1.4}$$

Zadeh voorziet tevens een transformatie van vaagverzamelingen naar klassieke verzamelingen door een ondergrens α op te leggen aan de lidmaatschapsgraad. Op deze manier wordt voor elke $\alpha \in [0, 1]$ de α-snede van een vaagverzameling $\widetilde{V}$ bepaald als:

$$\widetilde{V}_\alpha = \{x | x \in X \wedge \mu_{\widetilde{V}}(x) \geq \alpha\}. \tag{1.5}$$

Verder formuleert Zadeh veralgemeningen van enkele belangrijke operatoren uit de verzamelingenleer:

$$\text{(Gelijkheid)} \qquad \widetilde{V}_1 = \widetilde{V}_2 \Leftrightarrow \forall x \in X : \mu_{\widetilde{V}_1}(x) = \mu_{\widetilde{V}_2}(x) \tag{1.6}$$

$$\text{(Deelverzameling)} \qquad \widetilde{V}_1 \subseteq \widetilde{V}_2 \Leftrightarrow \forall x \in X : \mu_{\widetilde{V}_1}(x) \leq \mu_{\widetilde{V}_2}(x) \tag{1.7}$$

$$\text{(Complement)} \qquad \forall x \in X : \mu_{\mathrm{co}\left(\widetilde{V}\right)}(x) = 1 - \mu_{\widetilde{V}}(x) \tag{1.8}$$

$$\text{(Unie)} \quad \forall x \in X : \mu_{\widetilde{V}_1 \cup \widetilde{V}_2}(x) = \max\left(\mu_{\widetilde{V}_1}(x), \mu_{\widetilde{V}_2}(x)\right) \tag{1.9}$$

$$\text{(Doorsnede)} \quad \forall x \in X : \mu_{\widetilde{V}_1 \cap \widetilde{V}_2}(x) = \min\left(\mu_{\widetilde{V}_1}(x), \mu_{\widetilde{V}_2}(x)\right). \tag{1.10}$$

De kardinaliteit van een vaagverzameling is gedefinieerd als de som van alle lidmaatschapsgraden. De definities van doorsnede en unie zijn niet willekeurig gekozen. Enerzijds, naar analogie met de klassieke verzamelingenleer, is de unie van $\widetilde{V}_1$ en $\widetilde{V}_2$ de kleinste vaagverzameling waarvan zowel $\widetilde{V}_1$ als $\widetilde{V}_2$ een deelverzameling zijn. Duaal geldt ook dat de doorsnede van $\widetilde{V}_1$ en $\widetilde{V}_2$ de grootste vaagverzameling is, die nog een deelverzameling van zowel $\widetilde{V}_1$ als $\widetilde{V}_2$ is. Anderzijds geldt er ook dat:

$$\forall x \in X : x \in \left(\widetilde{V}_1 \cap \widetilde{V}_2\right)_\alpha \Leftrightarrow \min(\mu_{\widetilde{V}_1}(x), \mu_{\widetilde{V}_2}(x)) \geq \alpha. \tag{1.11}$$

Waar deze twee eigenschappen aangeven dat de doorsnede en unie zoals gedefinieerd door Zadeh uiterst intuïtief zijn, kan de belangrijke rol van deze twee operatoren ook meer formeel worden aangetoond. Het is namelijk zo dat de structuur $(\mathcal{F}(X), \cap, \cup, \mathrm{co}(.))$ voldoet aan alle wetten van een Boolese structuur, op twee na [7]. Deze twee wetten zijn de volgende:

$$\text{(Contradictie)} \quad \widetilde{V} \cap \mathrm{co}\left(\widetilde{V}\right) = \emptyset \tag{1.12}$$

$$\text{(Uitgesloten derde)} \quad \widetilde{V} \cup \mathrm{co}\left(\widetilde{V}\right) = X. \tag{1.13}$$

Deze wetten zijn wel partieel voldaan aangezien:

$$\forall \alpha \in]0.5, 1] \quad : \quad \left(\widetilde{V} \cap \mathrm{co}\left(\widetilde{V}\right)\right)_\alpha = \emptyset \tag{1.14}$$

$$\forall \alpha \in [0, 0.5[\quad : \quad \left(\widetilde{V} \cup \mathrm{co}\left(\widetilde{V}\right)\right)_\alpha = X. \tag{1.15}$$

In latere werken wordt het gebruik van minimum en maximum voor de implementatie van $\cap$ en $\cup$ verruimd tot triangulaire normen en triangulaire conormen [7] die als volgt worden gedefinieerd.

Definitie 1.3 (Triangulaire norm)
Een triangulaire norm is een functie $t : [0,1]^2 \to [0,1]$ *die voldoet aan de volgende eigenschappen:*

$$\textit{(Commutativiteit)} \qquad \forall (x,y) \in [0,1]^2 : t(x,y) = t(y,x) \tag{1.16}$$

$$\textit{(Associativiteit)} \qquad \forall (x,y,z) \in [0,1]^3 : t(x,t(y,z)) = t(t(x,y),z) \tag{1.17}$$

$$\textit{(Identiteitswet)} \qquad \forall x \in [0,1] : t(x,1) = x \tag{1.18}$$

$$\textit{(Monotoniteit)} \qquad \forall (x,y,z) \in [0,1]^3 : (y \leq z) \Rightarrow t(x,y) \leq t(x,z). \tag{1.19}$$

Definitie 1.4 (Triangulaire conorm)
Een triangulaire conorm is een functie $s : [0,1]^2 \to [0,1]$ *die voldoet aan de volgende eigenschappen:*

$$\textit{(Commutativiteit)} \qquad \forall (x,y) \in [0,1]^2 : s(x,y) = s(y,x) \tag{1.20}$$

$$\textit{(Associativiteit)} \qquad \forall (x,y,z) \in [0,1]^3 : s(x,s(y,z)) = s(s(x,y),z) \tag{1.21}$$

$$\textit{(Identiteitswet)} \qquad \forall x \in [0,1] : s(x,0) = x \tag{1.22}$$

$$\textit{(Monotoniteit)} \qquad \forall (x,y,z) \in [0,1]^3 : (y \leq z) \Rightarrow s(x,y) \leq s(x,z). \tag{1.23}$$

Beide klassen van operatoren zijn aan elkaar gekoppeld via de Wet van De Morgan die stelt dat:

$$t(x,y) = 1 - s\left(1-x, 1-y\right). \tag{1.24}$$

Elk triplet $(t, s, 1-.)$ waarvoor de Wet van De Morgan is voldaan, wordt een De Morgan triplet genoemd. Op basis van triangulaire normen en triangulaire conormen is het mogelijk de doorsnede en unie van vaagverzamelingen alternatief te gaan uitbreiden:

$$\forall x \in X : \mu_{\widetilde{V}_1 \cap_t \widetilde{V}_2}(x) = t\left(\mu_{\widetilde{V}_1}(x), \mu_{\widetilde{V}_2}(x)\right) \tag{1.25}$$

$$\forall x \in X : \mu_{\widetilde{V}_1 \cup_s \widetilde{V}_2}(x) = s\left(\mu_{\widetilde{V}_1}(x), \mu_{\widetilde{V}_2}(x)\right). \tag{1.26}$$

Het kan makkelijk worden aangetoond dat minimum en maximum respectievelijk de puntsgewijs grootste triangulaire norm en de puntsgewijs kleinste triangulaire conorm zijn en dat bovendien $(\min, \max, 1-.)$ een De Morgan triplet is. Dit betekent dat de operatoren $\cap_t$ en $\cup_s$ gereduceerd kunnen worden tot

Zadeh's operatoren $\cap$ en $\cup$ door gebruik van respectievelijk min en max. Twee voorbeelden van andere, veelgebruikte De Morgan triplets, zijn gebaseerd op het product en de probabilistische som:

$$t_P(x,y) = xy \tag{1.27}$$
$$s_P(x,y) = x+y-xy \tag{1.28}$$

en de Łukasiewics triangulaire norm en triangulaire conorm:

$$t_L(x,y) = \max(x+y-1,0) \tag{1.29}$$
$$s_L(x,y) = \min(x+y,1). \tag{1.30}$$

Dit laatste paar heeft twee interessante eigenschappen:

$$t_L(x,1-x) = \max(x+1-x-1,0) = 0 \tag{1.31}$$
$$s_L(x,1-x) = \min(x+1-x,1) = 1 \tag{1.32}$$

wat impliceert dat de structuur $(\mathcal{F}(X), \cap_{t_L}, \cup_{s_L}, \mathrm{co}(.))$ voldoet aan de contradictiewet en de wet van de uitgesloten derde. De genoemde triangulaire normen en triangulaire conormen zullen een belangrijke rol spelen in het vervolg van deze thesis.

Een fundamenteel aspect van de vaagverzamelingenleer is het uitbreidingsprincipe van Zadeh [6]. Dit is een methode om klassieke wiskundige relaties uit te breiden tot vaagrelaties:

Definitie 1.5 (Uitbreidingsprincipe van Zadeh)
Gegeven de verzamelingen X_1,...,X_n en Y en beschouw de relatie f zodat:

$$f : X_1 \times ... \times X_n \to \mathcal{P}(Y) \tag{1.33}$$

dan is de uitgebreide (Zadeh) relatie $\widetilde{f}$ van f bepaald als:

$$\begin{aligned} \widetilde{f} : \mathcal{F}(X_1) \times ... \times \mathcal{F}(X_n) &\to \mathcal{F}(Y) \\ \left(\widetilde{V}_1, ..., \widetilde{V}_n\right) &\mapsto \widetilde{f}\left(\widetilde{V}_1, ..., \widetilde{V}_n\right). \end{aligned} \tag{1.34}$$

Hierbij is $\widetilde{f}\left(\widetilde{V}_1, ..., \widetilde{V}_n\right)$ een vaagverzameling over Y waarbij voor alle $y \in Y$ geldt:

$$\mu_{\widetilde{f}(\widetilde{V}_1,...,\widetilde{V}_n)}(y) = \begin{cases} \sup_{\mathbf{x} \in V_{(f,y)}} \min\left(\mu_{\widetilde{V}_1}(\mathbf{x}_1), ..., \mu_{\widetilde{V}_n}(\mathbf{x}_n)\right) & \textbf{\textit{als}} \quad y \in \mathrm{im}(f) \\ 0 & \textbf{\textit{als}} \quad y \notin \mathrm{im}(f) \end{cases}$$

waarbij $\mathrm{im}(f)$ *het beeld of codomein van de relatie f is en $V_{(f,y)}$ de verzameling van n-tuples uit het domein van f die onder f afgebeeld worden op y.*

Dat het uitbreidingsprincipe van Zadeh meer is dan een intuïtieve notie van vervaging van afbeeldingen, is later aangetoond door De Cooman [8].

1.3 Multiverzamelingen

In het vervolg van deze thesis zal er regelmatig worden gesteund op het concept 'multiverzameling'. Naast een korte samenvatting van de bestaande literatuur rond dit concept, zullen we een aantal bijkomende definities invoeren die in het vervolg van deze thesis worden gebruikt. Indien verderop het concept van multiverzamelingen wordt gebruikt, verwijzen we naar deze sectie om de nodige definities te raadplegen.

Het concept 'multiverzameling' is een veralgemening van het concept 'verzameling', in de zin dat een multiverzameling een collectie is, waar een element meerdere keren kan in voorkomen. Hoewel de formalisering van dit concept lang op zich heeft laten wachten, is het principe reeds lang gekend. Zo stelt de fundamentele stelling van de algebra dat een veeltermvergelijking van de n^{de} graad, steeds n oplossingen heeft, waarbij identieke oplossingen kunnen bestaan. De oplossingen van een veeltermvergelijking kunnen dus niet adequaat worden voorgesteld als een gewone verzameling. Een ander voorbeeld vinden we bij relationele databanken, die theoretisch gestoeld zijn op verzamelingenleer, maar waarbij het praktische resultaat van een projectie-operatie kan resulteren in een multiverzameling, eerder dan een verzameling [9].

De term 'multiverzameling' als dusdanig, is voor het eerst geïntroduceerd door de Bruijn in persoonlijke communicatie met Knuth [10, 11]. In dezelfde periode introduceerde Yager een theorie van *bags*, een term die dezelfde lading dekt [9]. Een formalisering van het concept 'multiverzameling' is toe te schrijven aan Blizard [12, 13, 14, 15], die als eerste multiverzamelingen behandelt als een fundamenteel concept, eerder dan het af te leiden van verzamelingen. Voorts is het concept 'multiverzameling' onmiskenbaar verbonden met L-vaagverzamelingen, waarbij $L = \mathbb{N}$. In het kader van deze thesis zullen we de definitie van Yager gebruiken.

Definitie 1.6 (Multiverzameling)
Gegeven een universum X. Een multiverzameling M over het universum X wordt gekarakteriseerd door de multipliciteitsfunctie:

$$\omega_M : X \to \mathbb{N} \tag{1.35}$$

waarbij, voor $x \in X$, $\omega_M(x)$ de multipliciteit van x in M wordt genoemd. De verzameling van alle multiverzamelingen over X wordt genoteerd als $\mathcal{M}(X)$. De verzameling van alle multiverzamelingen over X met kardinaliteit n wordt genoteerd als $\mathcal{M}_n(X)$.

De multipliciteit van een element geeft aan hoeveel keer dat element voorkomt in de multiverzameling. Yager definieert in zijn werk de volgende drie operaties voor multiverzamelingen:

$$\text{(Unie)} \quad \forall x \in X : \omega_{A \cup B}(x) = \max\left(\omega_A(x), \omega_B(x)\right) \tag{1.36}$$

$$\text{(Doorsnede)} \quad \forall x \in X : \omega_{A \cap B}(x) = \min\left(\omega_A(x), \omega_B(x)\right) \tag{1.37}$$

$$\text{(Som)} \quad \forall x \in X : \omega_{A \oplus B}(x) = \omega_A(x) + \omega_B(x). \tag{1.38}$$

Voorts wordt de $\in$-operator voor multiverzamelingen als volgt gedefinieerd:

$$x \in M \Leftrightarrow \omega_M(x) > 0. \tag{1.39}$$

Voor $k \in \mathbb{N}$ is de k-snede van een multiverzameling M een gewone verzameling als volgt:

$$M_k = \{x | x \in X \wedge \omega_M(x) \geq k\}. \tag{1.40}$$

1.4 Possibiliteitstheorie

1.4.1 Definities

De possibiliteitstheorie is een theorie die onzekerheid behandelt in het kader van onvolledige informatie. De theorie is voor het eerst geïntroduceerd door Zadeh [16], die de term 'possibiliteit' hiervoor ontleende aan Gaines en Kohout [17]. Echter, in zijn initieel werk geeft Zadeh nog geen sluitende definitie voor wat possibiliteit precies is. Dit haalt Zadeh zelf aan: "Since our intuition concerning the properties of possibility distributions is not as yet well developed, some of the definitions which are formulated in the sequel should be viewed as provisional in nature." [16]. Door dit oorspronkelijke gebrek aan volwaardige interpretatie, heeft de ontwikkeling van de possibiliteitstheorie geen eenduidig verloop gekend. De belangrijkste werken omtrent de theorie worden hier besproken.

In het basiswerk over de possibiliteitstheorie ([16]) poogt Zadeh om een onzekerheidstheorie te construeren rond het door hem gedefinieerde concept 'vaagverzameling'. Meer specifiek vertrekt Zadeh vanuit de toekenning van waarden aan een variabele X waarvoor, op een linguïstische manier, een voorwaarde wordt gesteld. Laat bij wijze van voorbeeld de variabele X de leeftijd van Jan voorstellen en laat gegeven zijn dat Jan jong is. Volgens Zadeh wordt een dergelijke linguïstische voorwaarde onderbouwd door een vaagverzameling $\widetilde{V}$ (die model staat voor 'jong') en is de possibiliteit dat X gelijk is aan een bepaalde leeftijd x proportioneel aan de compatibiliteit van x met $\widetilde{V}$. Met andere woorden, de possibiliteit dat $X = x$ wordt gegeven door $\mu_{\widetilde{V}}(x)$. Op die manier kan volgens Zadeh een possibiliteitsverdeling worden afgeleid uit de lidmaatschapsfunctie van $\widetilde{V}$. Hiermee legt Zadeh een link tussen de vaagverzamelingenleer en de possibiliteitstheorie. Vanuit deze definitie legt Zadeh vervolgens een verband met een aantal concepten uit de waarschijnlijkheidsleer, hetgeen een aanzet geeft tot definitie van possibiliteitsmaten en conditionele possibiliteitsverdelingen. Zadeh linkt vaagverzamelingen aan possibiliteiten door te stellen dat de possibiliteitstheorie een theorie van linguïstische onzekerheid is, maar blijft aan de oppervlakte over de betekenis daarvan. Dit heeft als gevolg dat Zadeh géén gedetailleerde uiteenzetting geeft over de verschillen tussen de possibiliteitstheorie en de waarschijnlijkheidsleer.

Een meer stabiele poging om een theorie van possibiliteiten op te bouwen wordt toegeschreven aan Dubois en Prade [18]. Daar waar Zadeh vertrekt van

vaagverzamelingen en poogt hieraan een onzekerheidstheorie te koppelen, analyseren Dubois en Prade eerst en vooral het begrip 'onzekerheid'. Ze stellen dat onzekerheid kan worden veroorzaakt door twee orthogonale aspecten: variabiliteit enerzijds en onvolledigheid anderzijds. Variabiliteit houdt in dat er een zekere willekeur bestaat over de uitkomst van een experiment. Bijvoorbeeld bij het gooien van een (eerlijke) dobbelsteen bestaat onzekerheid over welke zijde van de dobbelsteen naar boven zal liggen na het gooien. Deze onzekerheid wordt echter enkel en alleen veroorzaakt door het feit dat er verschillende uitkomsten bestaan voor het experiment. Anders gezegd, de uitkomst van het experiment is variabel. Onvolledigheid houdt in dat er onvoldoende informatie voorhanden is om een antwoord te formuleren op een vraag. Wanneer er bijvoorbeeld gevraagd wordt naar de geboortedatum van Justine Henin, dan is het antwoord op die vraag intrinsiek niet variabel. Door een gebrek aan informatie kan het echter zijn dat er onzekerheid bestaat over de gevraagde geboortedatum. Deze onzekerheid is typisch omgekeerd evenredig met de informatie die voorhanden is. Wanneer we Justine Henin niet kennen, kunnen we geen uitspraak doen over haar geboortedatum. Echter, wanneer geweten is dat Justine Henin een Belgische vrouw is, dan weten we dat haar geboortedatum niet voor 1830 kan liggen. Het vrijgeven van informatie reduceert de onzekerheid over haar geboortedatum. Vertrekkende van dit onderscheid bestuderen Dubois en Prade vervolgens de waarschijnlijkheidsleer. De waarschijnlijkheidsleer vertrekt van een uitkomstenruimte Ω, een algebra van meetbare deelverzamelingen $\mathscr{A} \subseteq \mathcal{P}(\Omega)$ (de gebeurtenissen) en een waarschijnlijkheidsmaat $\mathcal{W} : \mathscr{A} \to [0, 1]$ die aan volgende eigenschappen voldoet:

$$\mathcal{W}(\emptyset) = 0 \tag{1.41}$$

$$\mathcal{W}(\Omega) = 1 \tag{1.42}$$

$$(A \cap B = \emptyset) \Rightarrow (\mathcal{W}(A \cup B) = \mathcal{W}(A) + \mathcal{W}(B)) . \tag{1.43}$$

Het triplet $(\Omega, \mathscr{A}, \mathcal{W})$ wordt een waarschijnlijkheidssruimte genoemd. Wanneer we ons voor de eenvoud beperken tot het discrete geval, dan kunnen we een waarschijnlijkheidsverdeling associëren aan Ω:

$$\forall \omega \in \Omega : \Pr(\omega) = \mathcal{W}(\{\omega\}) \tag{1.44}$$

zodat:

$$\sum_{\omega \in \Omega} \Pr(\omega) = 1. \tag{1.45}$$

Dubois en Prade onderscheiden drie interpretaties van deze waarschijnlijkheidsleer. De eerste interpretatie stelt dat de waarschijnlijkheid van een uitkomst gelijk is aan het aantal gebeurtenissen waarin deze voorkomt, gedeeld door het totaal aantal gebeurtenissen. Voorwaarde hierbij is dat alle uitkomsten even mogelijk zijn. De tweede interpretatie is de *frequentistische* interpretatie, waarbij waarschijnlijkheden worden beschouwd als de limiet van convergerende, gemeten frequenties. De eerste twee interpretaties zijn van toepassing in de context van herhaalbare experimenten, zoals bijvoorbeeld het gooien met

een dobbelsteen. Wanneer deze veronderstelling niet opgaat, beschouwt men de derde interpretatie, namelijk de *subjectivistische* interpretatie. Onder deze interpretatie worden waarschijnlijkheden beschouwd als graden van geloof dat een bepaalde singuliere (d.i. alleenstaande) gebeurtenis zich zal voordoen. Een voorbeeld van een singuliere gebeurtenis is de uitslag van verkiezingen. Dubois en Prade stellen vast dat er van deze drie interpretaties, slechts één interpretatie in staat is om onzekerheid door onvolledigheid te modelleren, namelijk de *subjectivistische* interpretatie. Echter, zij merken op dat het raamwerk met één enkele waarschijnlijkheidsverdeling, niet in staat is om onderscheid te maken tussen onzekerheid door variabiliteit en onzekerheid door onvolledigheid. Zo kan men niet afleiden uit een uniforme verdeling, waaróm deze verdeling uniform is. Is dit het resultaat van een groot aantal experimenten, of is dit een uitdrukking van een totaal gebrek aan informatie? Een bijkomend probleem met de klassieke aanpak, is het gebrek aan schaalbaarheid in functie van een variabele uitkomstenruimte Ω bij onzekerheid door onvolledigheid. Dit is aangetoond door Shafer [19] en kan worden ingezien met een eenvoudig voorbeeld.

Voorbeeld 1.1
Bij de vraag "Is Justine Henin geboren voor of na 1980?" geeft de subjectivistische *probabilist bij een totaal gebrek aan informatie als antwoord een verdeling* $\Pr_1(\textit{voor 1980}) = \Pr_1(\textit{na 1980}) = \frac{1}{2}$ *waarbij* $\Omega_1 = \{\textit{voor 1980}, \textit{na 1980}\}$. *Met hetzelfde gebrek aan informatie, zal het antwoord op de vraag "Is Justine Henin geboren voor 1970, na 1980 of daartussen?" gegeven zijn als* $\Pr_2(\textit{voor 1970}) = \Pr_2(\textit{na 1980}) = \Pr(\textit{tussen 1970 en 1980}) = \frac{1}{3}$ *met* $\Omega_2 = \{\textit{voor 1970}, \textit{na 1980}, \textit{tussen 1970 en 1980}\}$. *Echter, volgens de axioma's van de waarschijnlijkheidsleer zijn beide verdelingen in tegenspraak met elkaar, want aangezien:*

$$\{\textit{tussen 1970 en 1980}\} \cup \{\textit{voor 1970}\} = \{\textit{voor 1980}\} \tag{1.46}$$

en

$$\{\textit{tussen 1970 en 1980}\} \cap \{\textit{voor 1970}\} = \emptyset \tag{1.47}$$

zou er moeten gelden dat:

$$\Pr(\textit{voor 1980}) = \Pr(\textit{voor 1970}) + \Pr(\textit{tussen 1970 en 1980}) \tag{1.48}$$

hetgeen niet voldaan is.

In het algemeen is het zo dat het gebruik van uniforme verdelingen faalt bij het modelleren van het ontbreken van informatie, doordat het opsplitsen van een gebeurtenis A in meerdere disjuncte deelgebeurtenissen $\{A_i | i \in \{1, ..., n\}\}$ leidt tot een herverdeling van de totale probabiliteit, in plaats van een herverdeling van de probabiliteit van A. Vanuit dit standpunt redeneren Dubois en Prade dat er nood is aan een theorie waarin onzekerheid door onvolledigheid correct wordt behandeld. Het is precies deze nood die ze willen tegemoetkomen met de possibiliteitstheorie[1].

[1] De possibiliteitstheorie is een speciaal geval van de theorie van onnauwkeurige waarschijnlijkheden waarin variabiliteit en onvolledigheid samen worden behandeld [20]. Deze theorie valt buiten het bestek van deze thesis.

Gegeven een verzameling van uitkomsten Ω, beschouwen Dubois en Prade een vertrouwensmaat $g : 2^{\Omega} \to [0, 1]$ die voldoet aan de volgende eigenschappen:

$$\text{(Randvoorwaarde 1)} \qquad g(\emptyset) = 0 \tag{1.49}$$

$$\text{(Randvoorwaarde 2)} \qquad g(\Omega) = 1 \tag{1.50}$$

$$\text{(Monotoniteit)} \quad A \subseteq B \Rightarrow g(A) \leq g(B). \tag{1.51}$$

Deze vertrouwensmaat impliceert dat voor een gegeven $A \subseteq \Omega$ (een gebeurtenis) $g(A)$ het vertrouwen uitdrukt in het optreden van A. Aldus is $\emptyset$ de onmogelijke gebeurtenis en Ω de zekere gebeurtenis. Merk op dat als g voldoet aan het additiviteitsaxioma:

$$\forall A \subseteq \Omega : \forall B \subseteq \Omega : (A \cap B = \emptyset) \Rightarrow (g(A \cup B) = g(A) + g(B)) \tag{1.52}$$

dan is g een waarschijnlijkheidsmaat. Door de monotoniteit van g geldt er dat:

$$\begin{aligned} \forall A \subseteq \Omega : \forall B \subseteq \Omega \quad &: \quad g\left(A \cup B\right) \geq \max\left(g\left(A\right), g\left(B\right)\right) \\ \forall A \subseteq \Omega : \forall B \subseteq \Omega \quad &: \quad g\left(A \cap B\right) \leq \min\left(g\left(A\right), g\left(B\right)\right). \end{aligned}$$

In een extreem geval voor g krijgen we bijgevolg:

$$\forall A \subseteq \Omega : \forall B \subseteq \Omega : \Pi\left(A \cup B\right) = \max\left(\Pi\left(A\right), \Pi\left(B\right)\right) \tag{1.53}$$

hetgeen een possibiliteitsmaat wordt genoemd en in overeenstemming is met [16]. Dubois en Prade merken op dat Π onder deze definitie inderdaad een possibiliteitsmaat mag worden genoemd om drie redenen. Voor de eerste reden beschouwen ze een Boolese versie van de theorie. Gegeven een verzameling $E \subseteq \Omega$, die als zeker wordt beschouwd, dan is het mogelijk een functie Π_E te definiëren zodat:

$$\Pi_E(A) = \begin{cases} 1 & \textbf{als} \quad A \cap E \neq \emptyset \\ 0 & \textbf{als} \quad A \cap E = \emptyset. \end{cases} \tag{1.54}$$

Op deze manier geeft $\Pi_E(A)$ aan of A mogelijk is, of niet. Meer specifiek, A is mogelijk als A niet in tegenspraak is met E, hetgeen zeker is. De functie Π_E voldoet hierbij aan de voorwaarde $\Pi_E(A \cup B) = \max\left(\Pi_E(A), \Pi_E(B)\right)$. Als tweede reden merken Dubois en Prade op dat voor een gebeurtenis A en de tegenovergestelde gebeurtenis $\overline{A}$, minstens één van deze gebeurtenissen mogelijk moet zijn. Het kan worden ingezien dat er geldt:

$$\Pi\left(A \cup \overline{A}\right) = \Pi(\Omega) = 1. \tag{1.55}$$

Als derde reden wordt aangehaald dat de definitie overeenstemt met de 'fysieke' interpretatie van possibiliteit [18]. Met andere woorden, de mogelijkheid van gebeurtenis $A \cup B$ wordt bepaald door de mogelijkheid van de 'makkelijkste' van deze twee gebeurtenissen. In het kader van deze thesis zal steeds worden gesteld dat Ω een eindige verzameling is. In dat geval, kan Π volledig worden gekarakteriseerd door singleton gebeurtenissen, aangezien er geldt dat:

$$\forall A \subseteq \Omega : \Pi\left(A\right) = \sup_{a \in A} \Pi\left(\{a\}\right). \tag{1.56}$$

Dit motiveert het bestaan van een possibiliteitsverdeling:

$$\pi : \Omega \to [0,1] : \pi(\omega) = \Pi(\{\omega\}) \tag{1.57}$$

die naar analogie met de normeringsvoorwaarde voor waarschijnlijkheidsverdelingen (1.45) voldoet aan:

$$\sup_{\omega \in \Omega} \pi(\omega) = 1. \tag{1.58}$$

Het kan makkelijk worden aangetoond dat het probleem van schaalbaarheid in functie van een veranderende uitkomstenruimte Ω niet aanwezig is in het possibilistische raamwerk. Dit komt door wat men de maxitiviteit van possibiliteiten (zie (1.53)) noemt, in tegenstelling tot additiviteit (zie (1.52)) bij waarschijnlijkheden. De niet-additiviteit van possibiliteiten leidt tot de interessante observatie dat de possibiliteitstheorie een speciaal geval is. De afbeeldingsruimte $[0,1]$ waarin possibiliteiten worden uitgedrukt, kan namelijk worden veralgemeend tot een totaal geordende verzameling L, die zelfs ordinaal van aard kan zijn. Een belangrijk verschil met de waarschijnlijkheidsleer is dat possibiliteitsmaten niet zelf-duaal zijn. Dat wil zeggen dat $\Pi(A)$ niet noodzakelijk gelijk is aan $1 - \Pi(\overline{A})$. Interessant genoeg blijkt de duale maat van Π een vertrouwensmaat N te zijn waarvoor geldt:

$$\forall A \subseteq \Omega : \forall B \subseteq \Omega : \mathrm{N}(A \cap B) = \min(\mathrm{N}(A), \mathrm{N}(B)) \tag{1.59}$$

hetgeen eveneens een grens vormt voor een vertrouwensmaat g. Vertrouwensmaten die voldoen aan (1.59) worden necessiteitsmaten genoemd en er wordt gezegd dat ze minitief zijn. Analoog aan de redenering voor possibiliteitsmaten geven Dubois en Prade drie argumenten voor de naamgeving van necessiteitsmaten. Als opnieuw gegeven is dat $E \subseteq \Omega$ een zekere gebeurtenis is met:

$$\mathrm{N}_E(A) = \begin{cases} 1 & \textbf{als} \quad E \subseteq A \\ 0 & \textbf{anders} \end{cases} \tag{1.60}$$

dan geeft $\mathrm{N}_E(A)$ aan of A zeker is of niet. Meer specifiek, A is zeker als het een logisch gevolg is van een zekere gebeurtenis. Er geldt nu dat N_E voldoet aan (1.59). Voor necessiteiten vinden we dat $\min\left(\mathrm{N}(A), \mathrm{N}(\overline{A})\right) = 0$, wat overeenkomt met de intuïtie dat tegenstrijdige gebeurtenissen niet tegelijk zeker kunnen zijn. Een andere intuïtieve staving van deze concepten is:

$$\forall A \subseteq \Omega : \mathrm{N}(A) \leq \Pi(A) \tag{1.61}$$

hetgeen uitdrukt dat een gebeurtenis slechts maximaal zeker is in de mate dat die gebeurtenis mogelijk is. Dit wordt in de possibiliteitstheorie nog sterker uitgedrukt als:

$$\mathrm{N}(A) > 0 \Rightarrow \Pi(A) = 1 \tag{1.62}$$

en

$$\Pi(A) < 1 \Rightarrow \mathrm{N}(A) = 0. \tag{1.63}$$

Naast de interpretatie van Dubois en Prade zijn nog talloze andere interpretaties te vinden in de literatuur. Een zeer interessant werk is dit van de Britse econoom Shackle, die in 1961, lang voor Zadeh, reeds veel van de concepten van de possibiliteitstheorie heeft geïntroduceerd [21]. Meer bepaald vertrekt Shackle ook vanuit de observatie dat de waarschijnlijkheidsleer niet geschikt is om onzekerheid in bepaalde situaties voor te stellen. Eerst en vooral beschrijft Shackle wat hij noemt het rekenkundig obstakel van *distributionele onzekerheidsvariabelen*. Dit probleem houdt in dat, gegeven een vraag en een onvolledige verzameling van mogelijke antwoorden, het geen zin heeft een (waarschijnlijkheids)verdeling te construeren over de gegeven verzameling van antwoorden, aangezien niet geweten is wat de waarschijnlijkheid van de onbekende antwoorden is. Shackle stelt dat dit een gevolg is van de additiviteitsvoorwaarde van waarschijnlijkheden. Net als Dubois en Prade maakt Shackle een onderscheid tussen twee types vragen waarop een antwoord kan worden gegeven en deze types worden gedreven door een verschillende vorm van onzekerheid (variabiliteit versus onvolledigheid). Ook stelt Shackle, in het geval van onvolledigheid, dat de mate van mogelijkheid van een antwoord niet mag afhangen van het aantal alternatieve niet onmogelijke antwoorden. Dit komt in feite overeen met de kritiek van Dubois en Prade dat de waarschijnlijkheidsleer in het kader van onvolledigheid niet schaalbaar is.

Als oplossing voor dit probleem formuleert Shackle een informele theorie van '*mogelijke verrassing*'. Meer bepaald stelt Shackle dat de mate van onmogelijkheid van een gebeurtenis A gelijk is aan de mate van verrassing die veroorzaakt wordt wanneer gebeurtenis A optreedt. Verrassing veroorzaakt door een gebeurtenis komt dus overeen met necessiteit van de tegengestelde gebeurtenis. Anders gezegd, wanneer de *mogelijke verrassing* van een gebeurtenis A genoteerd wordt als $\mathrm{sur}(A)$, dan geldt:

$$\mathrm{sur}(A) = \mathrm{N}(\overline{A}) = 1 - \Pi(A). \tag{1.64}$$

Shackle stelt negen basisaxioma's voor in zijn theorie en leidt daar acht bijkomende stellingen uit af. Hoewel hij geen formele concepten zoals vertrouwensmaten aanhaalt, zijn er opvallende overeenkomsten tussen zijn axioma's en de principes van de possibiliteitstheorie. In het eerste axioma haalt Shackle aan dat kennis over een gebeurtenis A wordt bepaald door de combinatie van de *mogelijke verrassing* van A (d.i. $\mathrm{N}(\overline{A})$) en de *mogelijke verrassing* van $\overline{A}$ (d.i. $\mathrm{N}(A)$). We vinden in de axioma's van Shackle een equivalent van het maxitiviteitsprincipe en het normalisatieprincipe. Meer bepaald stelt Shackle dat:

$$\mathrm{sur}(A) = \inf_{B \subseteq A} \mathrm{sur}(B) \tag{1.65}$$

hetgeen inderdaad equivalent is met

$$\Pi(A) = \sup_{B \subseteq A} \Pi(B). \tag{1.66}$$

Voorts vinden we in het werk van Shackle ook dat, gegeven een universum X

in de zin van een zekere gebeurtenis, er moet gelden dat:

$$\inf_{x \in X} \text{sur}(\{x\}) = 0 \tag{1.67}$$

wat overeenstemt met het normalisatieprincipe. Daar waar de hiervoor besproken werken eerder een interpretatie (al dan niet met een wiskundige onderbouw) proberen te geven aan het concept van possibiliteit, worden de fundamenten van de possibiliteitstheorie uitvoerig bestudeerd en beschreven vanuit een puur formeel standpunt in [22]. Hoewel de volledige wiskundige formalisering van de possibiliteitstheorie buiten het bestek van deze thesis ligt, vermelden we [22] aangezien het aantoont dat de possibiliteitstheorie kan steunen op een goed onderbouwd raamwerk van vertrouwensmaten en vaagintegralen.

De possibiliteitstheorie zal in deze thesis een belangrijke rol spelen. Naast de wiskundige definities die hier worden ingevoerd, is de belangrijke conclusie dat we kunnen beschikken over possibiliteitsverdelingen die een voorstelling geven van een toestand van onvolledige informatie. Deze verdelingen kunnen worden geïnterpreteerd als de perceptie of de mening die een actor heeft. De possibiliteitstheorie is met andere woorden kennisbeschrijvend.

1.4.2 Conditionele possibiliteit

Een belangrijk aspect van de possibiliteitstheorie dat tot hiertoe nog niet is besproken, is dat van conditionele possibiliteit. De term 'conditioneel' behelst de kennis die men kan afleiden wanneer men (enkel en alleen) weet dat een bepaalde conditie waar is. De conditionele possibiliteit van een gebeurtenis A, gegeven een gebeurtenis C (conditie) wordt genoteerd als $\Pi(A|C)$ en wordt bepaald door de vergelijking van Hisdal [23]:

$$\Pi(A \cap C) = \min(\Pi(A|C), \Pi(C)). \tag{1.68}$$

Hisdal leidt hieruit af dat:

$$\Pi(A|C) \in \begin{cases} \{\Pi(A \cap C)\} & \textbf{als} \quad \Pi(C) > \Pi(A \cap C) \\ [\Pi(A \cap C), 1] & \textbf{als} \quad \Pi(C) = \Pi(A \cap C) \end{cases} \tag{1.69}$$

waaruit blijkt dat er niet altijd een unieke oplossing bestaat voor (1.68). Dubois en Prade stellen dat in het kader van de possibiliteitstheorie, de minst informatieve oplossing moet worden gekozen (principe van minimale specificiteit) en dus, gesteld dat $C \neq \emptyset$ en $A \neq \emptyset$ [24]:

$$\Pi(A|C) = \begin{cases} \Pi(A \cap C) & \textbf{als} \quad \Pi(C) > \Pi(A \cap C) \\ 1 & \textbf{als} \quad \Pi(C) = \Pi(A \cap C) \end{cases} \tag{1.70}$$

en dat $\Pi(A|C) = 1$ als $C = \emptyset$. Conditionele necessiteit kan worden bepaald door de dualiteit tussen possibiliteit en necessiteit:

$$\text{N}(A|C) = 1 - \Pi\left(\overline{A}|C\right). \tag{1.71}$$

Gegeven twee universa X en Y en de possibiliteitsverdelingen π_X en π_Y gedefinieerd over deze universa. Beschouw dan het universum $X \times Y$ en de possibiliteitsverdeling $\pi_{X\times Y}$. De verdelingen π_X en π_Y worden non-interactief ([6]) genoemd als er geldt dat:

$$\pi_{X\times Y}(x, y) = \min\left(\pi_X(x), \pi_Y(y)\right). \tag{1.72}$$

Het begrip 'afhankelijkheid' in de context van de possibiliteitstheorie wordt geformaliseerd door De Cooman, die in [22] stelt dat de verdelingen t-onafhankelijk zijn, met t een triangulaire norm, als:

$$\pi_{X\times Y}(x, y) = t\left(\pi_X(x), \pi_Y(y)\right). \tag{1.73}$$

Vermits de operator min de puntsgewijs grootste triangulaire norm is en dus bijgevolg de sterkste vorm van onafhankelijkheid representeert, zullen we zeggen dat min-onafhankelijke verdelingen onafhankelijk zijn.

1.4.3 Onzekerheid over Boolese proposities

Tot zover hebben we het gehad over onzekerheidstheorieën in het algemeen. Een interessante en reeds uitgebreid bestudeerde toepassing van zulke theorieën is het beschrijven van onzekerheid over klassieke Boolese proposities. Meer bepaald, in de klassieke Boolese logica worden proposities (uitspraken) beschouwd waarvan men op voorhand weet dat ze waar of vals zijn. Dit wordt het principe van bivalentie genoemd. Echter, in een context van onvolledige informatie is het mogelijk dat men de waarheidswaarde van een propositie niet kent. Dit probleem is voor het eerst bestudeerd door Kleene, die als oplossing een driewaardige logica voorstelt [25, 26]. Kleene beschouwt naast 'waar' en 'vals' een derde waarde, namelijk 'onbekend'. Vervolgens worden klassieke operatoren uit de Boolese logica uitgebreid naar de driewaardige logica. Een dergelijke logica vormt de basis voor de latere meerwaardige logica's [27]. Een belangrijke opmerking is dat de driewaardige logica van Kleene, net als de possibiliteitstheorie, kennisbeschrijvend is. De waarheidswaarden van Kleene geven een indicatie over wat men weet over proposities. Het kan geweten zijn dat een propositie waar is, dat ze vals is, of er kan helemaal niets geweten zijn. Dit wordt het principe van trivalentie genoemd.

Een interessante vraag is wat er gebeurt als we de kennisbeschrijvende possibiliteitstheorie gebruiken om kennis over Boolese proposities te beschrijven. Ook dit onderwerp is reeds uitgebreid bestudeerd. Het resultaat van dit onderzoek is een theorie van possibilistische waarheidswaarden, die een essentiële rol toebedeeld zijn in deze thesis. Een possibilistische waarheidswaarde is een possibiliteitsverdeling over het Boolese domein $\mathbb{B} = \{T, F\}$. Hierbij staat T voor waar en staat F voor vals. Dit leidt tot de volgende definitie [28, 29, 8].

Definitie 1.7 (Possibilistische waarheidswaarde)
Gegeven een Boolese propositie p. Een possibilistische waarheidswaarde is gedefinieerd als een possibiliteitsverdeling over $\mathbb{B} = \{T, F\}$:

$$\widetilde{p} = \left\{\left(T, \mu_{\widetilde{p}}(T)\right), \left(F, \mu_{\widetilde{p}}(F)\right)\right\} \tag{1.74}$$

zodat:

$$\mu_{\widetilde{p}}(T) = \pi_{\widetilde{p}}(T) \tag{1.75}$$
$$\mu_{\widetilde{p}}(F) = \pi_{\widetilde{p}}(F)\,. \tag{1.76}$$

De mogelijkheid dat p waar is wordt gegeven door $\mu_{\widetilde{p}}(T)$ en de mogelijkheid dat p vals is wordt gegeven door $\mu_{\widetilde{p}}(F)$.

Het domein van alle possibilistische waarheidswaarden wordt genoteerd als $\mathcal{F}(\mathbb{B})$. Een zeer interessante eigenschap in dit geval is de directe connectie tussen de possibiliteitsverdeling en de necessiteitsverdeling. Meer bepaald geldt er dat:

$$\mathrm{N}(\{F\}) = 1 - \mu_{\widetilde{p}}(T) \tag{1.77}$$
$$\mathrm{N}(\{T\}) = 1 - \mu_{\widetilde{p}}(F). \tag{1.78}$$

In volgende hoofdstukken zullen we gebruik maken van de notaties Pos en Nec om respectievelijk possibiliteiten en necessiteiten aan te duiden als volgt:

$$\mathrm{Pos}(p = T) \stackrel{\triangle}{=} 1 - \mathrm{Nec}(p = F) \stackrel{\triangle}{=} \mu_{\widetilde{p}}(T) \tag{1.79}$$
$$\mathrm{Pos}(p = F) \stackrel{\triangle}{=} 1 - \mathrm{Nec}(p = T) \stackrel{\triangle}{=} \mu_{\widetilde{p}}(F). \tag{1.80}$$

De operatoren Pos en Nec evalueren dus respectievelijk de mogelijkheid en de zekerheid van de toewijzing van een bepaalde Boolese waarde aan een propositie p. Door gebruik te maken van het uitbreidingsprincipe van Zadeh, is het mogelijk een uitgebreide definitie te geven voor de klassieke negatie-operator.

Definitie 1.8 (Zadeh-uitbreiding van $\neg$)
Gegeven een possibilistische waarheidswaarde $\widetilde{p}$, dan is de Zadeh-uitbreiding van $\neg$ gedefinieerd als:

$$\widetilde{\neg} : \mathcal{F}(\mathbb{B}) \to \mathcal{F}(\mathbb{B}) : \widetilde{p} \mapsto \widetilde{\neg}\,\widetilde{p} \tag{1.81}$$

waarbij:

$$\mu_{(\widetilde{\neg}\,\widetilde{p})}(T) = \mu_{\widetilde{p}}(F) \tag{1.82}$$
$$\mu_{(\widetilde{\neg}\,\widetilde{p})}(F) = \mu_{\widetilde{p}}(T). \tag{1.83}$$

Voor de Zadeh-uitbreiding van $\neg$ geldt er dat:

$$\forall \widetilde{p} \in \mathcal{F}(\mathbb{B}) : \widetilde{\neg}\,(\widetilde{\neg}\,\widetilde{p}) = \widetilde{p}. \tag{1.84}$$

In wat volgt zullen we de koppelnotatie voor possibilistische waarheidswaarden gebruiken. Onder deze notatie wordt een possibilistische waarheidswaarde voorgesteld als een koppel van lidmaatschapsgraden. Een possibilistische waarheidswaarde $\widetilde{p}$ wordt dus in koppelnotatie geschreven als:

$$(\mu_{\widetilde{p}}(T), \mu_{\widetilde{p}}(F))\,. \tag{1.85}$$

Op het domein $\mathcal{F}(\mathbb{B})$ worden volgende orderelaties gedefinieerd:

$$\widetilde{p} \geq \widetilde{q} \Leftrightarrow \begin{cases} \mu_{\widetilde{p}}(F) \leq \mu_{\widetilde{q}}(F) & \textbf{als} \quad \mu_{\widetilde{p}}(T) = \mu_{\widetilde{q}}(T) = 1 \\ \mu_{\widetilde{q}}(T) \leq \mu_{\widetilde{p}}(T) & \textbf{anders} \end{cases} \tag{1.86}$$

en

$$\widetilde{p} \leq \widetilde{q} \Leftrightarrow \begin{cases} \mu_{\widetilde{p}}(F) \geq \mu_{\widetilde{q}}(F) & \textbf{als} \quad \mu_{\widetilde{p}}(T) = \mu_{\widetilde{q}}(T) = 1 \\ \mu_{\widetilde{q}}(T) \geq \mu_{\widetilde{p}}(T) & \textbf{anders}. \end{cases} \tag{1.87}$$

Op eenzelfde manier kan ook de gelijkheidsrelatie worden gedefinieerd:

$$\widetilde{p} = \widetilde{q} \Leftrightarrow (\mu_{\widetilde{p}}(F) = \mu_{\widetilde{q}}(F) \wedge \mu_{\widetilde{p}}(T) = \mu_{\widetilde{q}}(T)) \,. \tag{1.88}$$

Laat ons tot slot van dit hoofdstuk enkele bijzondere possibilistische waarheidswaarden bespreken. In het geval waarbij $\widetilde{p} = (1, 0)$ betekent dit dat de waarheidswaarde van p zeker waar is. Andersom, in het geval waarbij $\widetilde{p} = (0, 1)$ betekent dit dat de waarheidswaarde van p zeker vals is. In het geval waarbij $\widetilde{p} = (1, 1)$ is er geen kennis voorhanden over de waarheidswaarde van p. Het is mogelijk dat p waar is, maar het is even mogelijk dat p vals is. De waarheidswaarde van p is dus onbekend.

1.5 Conclusie

In dit hoofdstuk is een overzicht gegeven van de belangrijkste literatuur waarop deze thesis is gestoeld. Hierbij zijn twee veralgemeningen van verzamelingen besproken: vaagverzamelingen en multiverzamelingen. Voorts is een bespreking gegeven van de possibiliteitstheorie. Dit is een onzekerheidstheorie waarbij er expliciet wordt verondersteld dat onzekerheid wordt veroorzaakt door onvolledige informatie. Wanneer de possibiliteitstheorie wordt gebruikt om onzekerheid over de waarheidswaarde van Boolese proposities te modelleren, ontstaat er een raamwerk van possibilistische waarheidswaarden. Dit concept zal een voorname rol spelen in het vervolg van deze thesis.

Hoofdstuk 2

Objecten en evaluatoren

2.1 Inleiding

In de voorbeschouwing van deze thesis is reeds een korte kadering voor dit werk gegeven binnen de hedendaagse informatiemaatschappij. Meer bepaald handelt deze thesis over de problematiek die ontstaat bij het beheer en de verwerking van zeer grote hoeveelheden data. In dat opzicht wordt er een subtiel maar reëel verschil gehanteerd tussen data en informatie. De term 'data' slaat in de context van deze thesis steeds op gegeven feiten en metingen. Meer specifiek bedoelen we met data gegevens die, al dan niet digitaal, opgeslagen of geregistreerd zijn. Wanneer data zo worden opgeslagen dat ze op een later tijdstip opnieuw kunnen worden geraadpleegd, dan zegt men dat de data persistent opgeslagen zijn. Een persistente collectie van data wordt een databank genoemd. De term 'informatie' handelt op een hoger niveau over de interpretatie die aan data wordt gegeven. Het onderscheid tussen data en informatie speelt een belangrijke rol in het beheer van data, in de zin dat een databank kan verzekeren dat de aanwezige data consistent is, maar nooit dat de aanwezige data correct is. Daar waar consistentie op een logische manier kan worden afgedwongen op basis van feiten en integriteitsregels, is de correctheid van data eerder gekoppeld aan de interpretatie ervan. De controle op correctheid vereist bijgevolg een zekere subjectieve connotatie[1]. In een ziekenhuis bijvoorbeeld worden metingen bij patiënten zoals lichaamstemperatuur en bloeddruk bijgehouden. Deze metingen vormen de data en artsen kunnen op basis van deze data informatie verschaffen over de toestand van patiënten. Bij het opslaan van dergelijke data kunnen consistentieregels worden geformuleerd, die stellen dat de lichaamstemperatuur binnen een bereik van 34°C en 43°C moet liggen. Echter, er kan geen enkele regel worden opgelegd die controleert of een waarde van 37.5°C voor patiënt x inderdaad overeenstemt met de werkelijke lichaamstemperatuur van deze patiënt. Het onderscheid tussen data en informatie wordt weerspiegeld in

[1] Motro en Rakov hebben maten van correctheid gedefinieerd, maar ze stellen dat deze maten minstens gedeeltelijk manueel berekend moeten worden [30].

het onderscheid dat wordt gemaakt tussen objecten en entiteiten.

In dit hoofdstuk zal in Sectie 2.2 eerst het onderscheid tussen objecten en entiteiten worden toegelicht en zal de problematiek van coreferente objecten worden besproken. Vervolgens wordt in Sectie 2.3 aangetoond waarom het zoeken naar coreferente objecten geen triviaal probleem is. Meer bepaald zal worden bekeken hoe onzekerheid sluipt in de beslissing aangaande de coreferentie van twee objecten. In Sectie 2.4 wordt een oplossing voor het probleem van coreferentiebepaling voorgesteld vanuit een possibilistisch standpunt. In het uitwerken van deze oplossing zal worden gesteund op een specifiek geval van de possibiliteitstheorie, met name possibilistische waarheidswaarden. De basisoperator van het possibilistische raamwerk voor coreferentiebepaling wordt een evaluator genoemd. Het zal blijken dat de eigenschappen waaraan een dergelijke operator moet voldoen, sterk afhangen van de veronderstellingen die worden gemaakt over de probleemstelling. In het bijzonder wordt bestudeerd hoe de eigenschappen van de coreferentierelatie al dan niet worden overgedragen op een evaluator. In Sectie 2.5 wordt het probleem van moeilijk-meetbaarheid van eigenschappen besproken en in Sectie 2.6 wordt de klasse van semantische evaluatoren besproken. In Sectie 2.7 wordt de problematiek van gedeeltelijke coreferentie besproken en het wordt aangetoond hoe dit probleem gerelateerd is aan het probleem van coreferentie. Sectie 2.8 biedt een overzicht van de belangrijkste bevindingen van dit hoofdstuk.

2.2 Objecten en entiteiten

Een object wordt in het kader van deze thesis als een axiomatisch beginsel ondersteld en is als volgt gedefinieerd.

Definitie 2.1 (Object)
Een object is een eenheid van data.

Om de algemeenheid te behouden worden er geen verdere vereisten geformuleerd in de definitie van een object. Het wordt dan ook benadrukt dat de term 'object' los staat van de interpretatie die er aan gegeven wordt binnen de informatica en in het bijzonder in de context van object-oriëntatie [31]. Omgekeerd is het wel zo dat de objectstandaard uit de informatica [31] een geldig voorbeeld is van wat in deze thesis met een object wordt bedoeld. Een willekeurig object zal steeds worden genoteerd als o, eventueel voorzien van een index om meerdere objecten tegelijk te benoemen. Het domein van een willekeurig object wordt genoteerd als O en wordt het objectuniversum of de objectruimte genoemd. Als een object bestempeld wordt als data, dan moet er een equivalent zijn voor informatie om een correcte redering te kunnen opbouwen. Een dergelijk equivalent is gegeven door de entiteiten uit de reële wereld die beschreven worden door objecten. Dit formaliseren we door de definitie van objectreferentie.

Definitie 2.2 (Objectreferentie)
Gegeven een universum O, dan bestaat er een universum van entiteiten $\mathcal{E}$ en een surjectieve, niet-injectieve referentiefunctie $\rho : O \to \mathcal{E}$ zodat voor alle $o \in O$, $\rho(o)$ de entiteit is waarnaar o refereert.

Objecten kunnen altijd logisch gegroepeerd worden volgens het type van entiteiten dat ze beschrijven. Dat wil zeggen dat het universum $\mathcal{E}$ gepartitioneerd kan worden in equivalentieklassen, zodat elke equivalentieklasse alle entiteiten van één type bevat. Met elk type zullen we een uniek label associëren.

Definitie 2.3 (Labelfunctie)
Voor een objectuniversum O wordt een labelfunctie gedefinieerd als:

$$\text{lab} : O \to \mathcal{L} : o \mapsto \text{lab}(o) \tag{2.1}$$

Objecten met eenzelfde label beschrijven entiteiten van eenzelfde type.

Voorbeeld 2.1
Laat ons deze definities verduidelijken met een voorbeeld. Veronderstel het object $o = 37.5$ en stel dat lab(o) =*'lichaamstemperatuur', dan weten we dat o een lichaamstemperatuur beschrijft. Merk op dat de eenheid waarin lichaamstemperatuur hier wordt gespecificeerd, niet gekend is. We zullen de eenheid waarin objecten een bepaald entiteitstype beschrijven als vast en gekend beschouwen. In het voorbeeld wil dit zeggen dat we weten dat $o = 37.5$ een lichaamstemperatuur beschrijft, gemeten in (bijvoorbeeld) graden Celsius.*

Afhankelijk van de structuur van het objectuniversum wordt een fundamenteel onderscheid gemaakt tussen atomaire en complexe objecten.

Definitie 2.4 (Atomair object)
Een atomair universum is een universum dat niet kan worden geschreven als een relatie over andere universa dan zichzelf. Objecten behorend tot een atomair universum worden atomaire objecten genoemd.

Voorbeelden van atomaire universa zijn de verzameling van de natuurlijke getallen $\mathbb{N}$ en de machtsverzameling van $\mathbb{N}$, genoteerd als $\mathcal{P}(\mathbb{N})$.

Definitie 2.5 (Complex object)
Een complex universum is een n-voudige relatie over n atomaire universa. Objecten behorend tot een complex universum worden complexe objecten genoemd.

De notatie O zal worden gebruikt om een willekeurig (complex) universum aan te duiden. Een willekeurig atomair universum wordt genoteerd als U en een willekeurig atomair object wordt genoteerd als u. Onder deze notatie kan een complex universum O worden geschreven als:

$$O = U_1 \times ... \times U_n. \tag{2.2}$$

Voorbeeld 2.2
In geografische toepassingen worden kaarten vaak uitgerust met points of interest *(POIs) om bezienswaardigheden aan te duiden. Hierbij worden POIs beschreven aan de hand van een lengtegraad, breedtegraad, naam en type. Elke POI wordt bijgevolg beschreven door een complex object behorende tot het universum:*

$$O_{POI} = U_1 \times U_2 \times U_3 \times U_4 \tag{2.3}$$

waarbij:

$$\begin{aligned} U_1 &= [-180, 180] \\ U_2 &= [-90, 90] \\ U_3 &= \mathcal{S} \\ U_4 &= \mathbb{T}_{POI}. \end{aligned}$$

Hierbij is $\mathcal{S}$ de verzameling van alle karakterstrings en $\mathbb{T}_{POI}$ de (aftelbare) verzameling van alle POI-types. De labelfunctie is gespecificeerd als volgt:

$$\begin{aligned} \forall u \in U_1 : \text{lab}(u) &= \textit{lengtegraad} \\ \forall u \in U_2 : \text{lab}(u) &= \textit{breedtegraad} \\ \forall u \in U_3 : \text{lab}(u) &= \textit{naam} \\ \forall u \in U_4 : \text{lab}(u) &= \textit{type} \\ \forall o \in O : \text{lab}(o) &= \textit{POI}. \end{aligned}$$

Definitie 2.6 (Deeluniversum)
Gegeven een objectuniversum $O = U_1 \times ... \times U_n$. Het universum U_i wordt het i^{de} deeluniversum van O genoemd.

Het verband tussen een complex object o en de atomaire objecten waaruit o is opgebouwd wordt gegeven door de projectie-operator voor complexe objecten.

Definitie 2.7 (Projectie van complexe objecten)
Gegeven een objectuniversum $O = U_1 \times ... \times U_n$. Voor een willekeurig object $o = (u_1, ..., u_n)$ wordt de i^{de} projectie gedefinieerd als volgt:

$$\forall i \in \{1, ..., n\} : \text{proj}_i(o) = u_i. \tag{2.4}$$

Het object $\text{proj}_i(o)$ wordt het i^{de} deelobject van o genoemd.

Definitie 2.8 (Eigenschap van een entiteit)
Gegeven een objectuniversum $O = U_1 \times ... \times U_n$ en een willekeurig object $o = (u_1, ..., u_n)$. Als $\rho(o) = e$, dan wordt de entiteit e_i beschreven door u_i de i^{de} eigenschap van e genoemd.

De structuur van een complex object toont een onmiddellijk verband met de definitie van het relationele databankmodel van Codd [1], waarbij tuples in een databank worden gedefinieerd als een n-voudige relatie. Het wordt echter benadrukt dat onze definitie van (complexe) objecten losstaat van belangrijke standaarden uit de informatica ([1, 31, 32]), maar eerder de compatibiliteit met deze standaarden nastreeft via een weinig specifieke, maar algemene definitie. Om deze compatibiliteit te verwezenlijken is het mogelijk om een universum O te voorzien van bijkomende structuren. In vele hedendaagse toepassingen, zoals bijvoorbeeld de *opmaaktaal* XML, worden complexe objecten bijvoorbeeld voorgesteld met behulp van een boomstructuur. In dit model is elke bladknoop equivalent met een atomair universum en is elke interne knoop equivalent met een semantische samenhangende groep van atomaire universa. Een dergelijke structuur kan worden in rekening gebracht door de definitie van een groeperingsfunctie.

Definitie 2.9 (Groeperingsfunctie)
Gegeven een objectuniversum $O = U_1 \times ... \times U_n$. Een groeperingsfunctie voor O is een functie:

$$\lambda : \mathcal{P}(\{\text{lab}(U_1), ..., \text{lab}(U_n)\}) \to \{0, 1\} \tag{2.5}$$

die voldoet aan:

$$\text{(1)} \qquad \forall i \in \{1, ..., n\} : \lambda(\{\text{lab}(U_i)\}) = 1 \tag{2.6}$$

$$\text{(2)} \qquad \lambda(\{\text{lab}(U_1), ..., \text{lab}(U_n)\}) = 1 \tag{2.7}$$

$$\text{(3)} \qquad (\lambda(A) = \lambda(B) = 1) \Rightarrow ((A \subseteq B) \vee (B \subseteq A) \vee (A \cap B = \emptyset)) \,. \tag{2.8}$$

Hierbij stelt $\mathcal{P}(\{\text{lab}(U_1), ..., \text{lab}(U_n)\})$ de machtsverzameling van de labels van de atomaire universa voor. Voor elke $A \in \mathcal{P}(\{\text{lab}(U_1), ..., \text{lab}(U_n)\})$ geeft $\lambda(A)$ aan of een verzameling van labels A, een semantisch samenhangende groep vormt (1) of niet (0). De voorwaarden vereisen dat labels van atomaire universa op zich een semantisch samenhangende groep zijn (1), dat de verzameling van alle labels een semantisch samenhangende groep is (2) en dat voldaan is aan de ouder-kind relatie van bomen(3).

Voorbeeld 2.3
Aansluitend bij Voorbeeld 2.2 kunnen we een groeperingsfunctie als volgt definiëren:

$$\begin{aligned}
\lambda(\{\textit{lengtegraad}\}) &= 1 \\
\lambda(\{\textit{breedtegraad}\}) &= 1 \\
\lambda(\{\textit{naam}\}) &= 1 \\
\lambda(\{\textit{type}\}) &= 1 \\
\lambda(\{\textit{lengtegraad}, \textit{breedtegraad}\}) &= 1 \\
\lambda(\{\textit{lengtegraad}, \textit{breedtegraad}, \textit{naam}, \textit{type}\}) &= 1.
\end{aligned}$$

Het probleem dat in deze thesis wordt bestudeerd, is gerelateerd aan de referentiefunctie ρ. Een belangrijke eigenschap van deze functie is haar surjectiviteit, hetgeen betekent dat:

$$\forall e \in \mathcal{E} : \exists o \in O : \rho(o) = e. \tag{2.9}$$

Voor elke entiteit e bestaat er dus minstens één object o dat e beschrijft. De functie ρ is echter niet-injectief. Het is bijgevolg perfect mogelijk dat twee verschillende objecten o_1 en o_2 dezelfde entiteit beschrijven, om verscheidene redenen die verder worden besproken. Zulke objecten worden coreferent genoemd: ze refereren naar eenzelfde entiteit.

Definitie 2.10 (Coreferentie van objecten)
Gegeven een universum O en een referentiefunctie ρ. Twee objecten $(o_1, o_2) \in O^2$ zijn coreferent, genoteerd als $o_1 \leftrightarrow o_2$ als:

$$\rho(o_1) = \rho(o_2). \tag{2.10}$$

Definitie 2.10 heeft als gevolg dat $\leftrightarrow$ een equivalentierelatie is en dus voldoet aan de volgende eigenschappen:

$$(1) \quad \forall o \in O \quad : o \leftrightarrow o \tag{2.11}$$

$$(2) \quad \forall (o_1, o_2) \in O^2 \quad : o_1 \leftrightarrow o_2 \Leftrightarrow o_2 \leftrightarrow o_1 \tag{2.12}$$

$$(3) \quad \forall (o_1, o_2, o_3) \in O^3 \quad : o_1 \leftrightarrow o_2 \wedge o_2 \leftrightarrow o_3 \Leftrightarrow o_1 \leftrightarrow o_3. \tag{2.13}$$

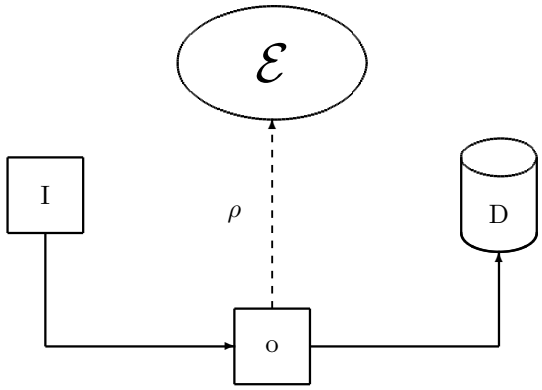
Deze eigenschappen worden respectievelijk reflexiviteit (1), symmetrie (2) en transitiviteit (3) genoemd. De overheveling van deze eigenschappen van de gelijkheidsrelatie in het domein van entiteiten naar de coreferentierelatie in het domein van objecten, is een belangrijk aandachtspunt in de ontwikkeling van een theoretisch raamwerk voor de behandeling van coreferentie en zal in de volgende hoofdstukken nog uitvoerig worden behandeld. Gegeven een collectie van (complexe) objecten, dan is het coreferentieprobleem bepaald als het vinden van alle koppels van coreferente objecten.

Voorbeeld 2.4
Beschouw aansluitend bij Voorbeelden 2.2 en 2.3 de volgende objecten. De oplossing voor het coreferentieprobleem is hier de verzameling van koppels $\{(o_1, o_2), (o_3, o_4)\}$, waarbij i voor o_i verwijst naar de waarde van 'eid'.

eid	**lengtegraad**	**breedtegraad**	**naam**	**type**
1	3.72288	51.038235	S.M.A.K	Museum
2	3.72327	51.038128	Stedelijk Museum voor Actuele Kunst	POI
3	3.72170	51.051970	Bakkerij Bloch	POI
4	3.721852	51.052233	Bloch	Bakkerij

Het coreferentieprobleem kan zich in veel verschillende contexten manifesteren. Voorbeeld 2.4 is een illustratie van het coreferentieprobleem binnen één (geografische) databank waarin data zijn opgeslagen. Het zoeken naar coreferente objecten is hierbij van belang voor enerzijds efficiënte opslag van data maar anderzijds ook voor een juiste berekening van statistieken. Het zoeken naar coreferente objecten binnen één databank wordt ook wel het opkuisen van een databank genoemd. Alternatief kan het coreferentieprobleem zich manifesteren door verschillende databanken te beschouwen. Een relevant voorbeeld hiervan is het samenvoegen van databanken tot een *data warehouse* met als doel de efficiënte (historische) analyse van data. Opnieuw is het belangrijk dat dubbele data worden geïdentificeerd om efficiënte opslag en juistheid van analyses te garanderen. Een ander relevant voorbeeld is het identificeren van personen op basis van biometrische gegevens. Het coreferentieprobleem wordt een heel stuk moeilijker wanneer de verschillende databanken niet rechtstreeks raadpleegbaar zijn. Stel bijvoorbeeld dat twee websites gedeeltelijk dezelfde producten te koop aanbieden, dan kan het van concurrentieel belang zijn een vergelijking te maken tussen beide portefeuilles. Hiervoor kunnen beide partijen wel elkaars website raadplegen, maar kunnen ze niet rechtstreeks de databanken aanspreken die schuilgaan achter de websites. Nog moeilijker wordt het wanneer objecten worden beschouwd in een context van multimedia, waarbij losse tekst, geluidsbestanden, beelden en videofragmenten allemaal beschrijvingen van entiteiten zijn. Een interessant voorbeeld is YouTube, een website waarop videobestanden kunnen worden geüpload en bekeken. Heel wat van deze bestanden betreffen coreferente bestanden. Gelet op het feit dat het 412 jaar zou duren om alle videobestanden op YouTube te bekijken en gelet op het feit dat er per minuut 13 uur aan videofragementen op YouTube bijkomen zou het ontdubbelen van YouTube een gigantische taak zijn die onmogelijk handmatig kan worden opgelost.

Alvorens een concrete oplossing voor het algemene corefentieprobleem te geven, is het nodig om een aantal bijzondere aspecten van dit probleem te bespreken. Een belangrijke vraag die eerst en vooral moet worden beantwoord, is de volgende. Waarom bestaat het coreferentieprobleem? Dit lijkt misschien een vreemde vraag, maar enige toelichting brengt duidelijkheid. Wanneer we kijken naar de geschiedenis van databeheer, is het ontwerp van databanken er steeds op gericht om anomalieën in data te vermijden. Een uiterst relevant voorbeeld dat we hierbij aanhalen, is het concept van normalisatie bij databanken, dat als doel heeft om opgeslagen data op een consistente manier aan te passen en te bewerken, zodat dubbele gegevensopslag daarbij wordt vermeden. Men kan zich dus afvragen of er een databank kan worden geconstrueerd die op elk willekeurig tijdstip vrij is van coreferente objecten, los van de coreferentie van objecten met zichzelf. Om deze vraag te beantwoorden maken we een abstractie van het invoerproces van objecten in databanken. Figuur 2.1 toont een databank D die objecten bevat uit het universum O. Deze objecten beschrijven entiteiten uit een universum $\mathcal{E}$ dat de reële wereld voorstelt. Een invoerproces I genereert objecten o die refereren naar entiteiten uit $\mathcal{E}$. Be-



Figuur 2.1: Invoer van data in een databron

langrijk hierbij is dat de objectreferentie vanuit het standpunt van I niet kan worden opgeslagen in D. Het begrip 'referentie' situeert zich op het niveau van informatie in plaats van het niveau van data. Het verband tussen D en $\mathcal{E}$ kan dus niet worden voorgesteld op niveau van data en het is precies dit verband dat we proberen te reconstrueren. Vanuit dit standpunt bekeken, kunnen we de eerder gestelde vraag herformuleren in termen van de databank (D) en het invoerproces (I). De vraag luidt dan: "Kunnen I en D zodanig worden bepaald dat de reconstructie van ρ triviaal wordt". Het antwoord is neen, tenzij in het triviale geval waarbij D slechts één object mag bevatten. Immers, I kan per definitie verschillende objecten genereren. Vermits er op het niveau van D geen enkele notie bestaat over de correctheid die I hanteert bij het refereren naar entiteiten, kan het nooit worden vermeden dat er onzekerheid bestaat over de reconstructie van ρ. Desondanks bestaan er regels van goed ontwerp om de onzekerheid omtrent de reconstructie van ρ tot een minimum te beperken. We halen drie van deze regels aan bij wijze van voorbeeld.

Een eerste regel benut het feit dat het universum O in heel wat situaties eindig is, of dat het minstens zinvol is om eindigheid te veronderstellen. Voorbeelden hiervan zijn geboortedata, straatnamen, postcodes, namen van steden ... Wanneer deze veronderstelling opgaat, wil dit zeggen dat D een exhaustieve lijst kan opstellen, waaruit I een keuze moet maken, zodat I geen onverwachte input kan doorgeven aan D (bijvoorbeeld het invullen van een voornaam wanneer een straatnaam wordt verwacht). Het maken van een keuze uit een lijst ondersteunt ook de stelling dat I zijn keuze doordachter maakt dan wanneer hij willekeurig informatie zou genereren (bijvoorbeeld omdat I niet weet wat er precies wordt verwacht). Deze techniek kan worden gezien als een eenduidige communicatie tussen I en D en wordt in vele informatiesystemen (vooral webtoepassingen) reeds toegepast.

Een tweede regel handelt over de granulariteit van objecten. In Definitie 2.2 wordt objectreferentie formeel gedefinieerd door middel van een functie. Dit houdt in dat er geen twee verschillende entiteiten e_1 en e_2 kunnen bestaan zodat, voor willekeurige $o \in O$, $\rho(o) = e_1$ en $\rho(o) = e_2$. Met andere woorden, een object kan niet refereren naar twee verschillende entiteiten. Echter, deze veronderstelling kan in de praktijk vervallen wanneer het ontwerp van D een ontoereikende granulariteit hanteert. Laat ons dit illustreren met een voorbeeld. Stel dat een bedrijf een klantenbestand wil bijhouden. Daartoe construeert de IT-afdeling een tabel met daarin het unieke klantennummer, de naam en de voornaam van de klant. Het is duidelijk dat de kans op invoer van coreferente klanten bijzonder hoog is, aangezien de uniciteit van naam en voornaam met betrekking tot het entiteitstype persoon zeker niet gegarandeerd is. Naast de hoge kans op coreferente objecten wordt het ook onmogelijk om coreferente objecten te zoeken, vermits gelijkheid van objecten geen zekerheid biedt op coreferentie van objecten. Immers, voor twee klanten met gelijke naam en voornaam kan niet met zekerheid worden gesteld dat het coreferente klanten betreft. Dit is in contradictie met de voorwaarde dat ρ een functie is. Als richtlijn voor ontwerp kan worden gesteld dat een object een entiteit in die mate moet beschrijven, dat de mogelijkheid van niet-coreferentie van gelijke objecten verwaarloosbaar klein is. Anders gezegd moet de granulariteit van het universum O (objecten) overeenstemmen met die van het universum $\mathcal{E}$ (entiteiten).

Een derde belangrijke regel motiveert het gebruik van standaardnotaties voor data. Deze regel houdt in dat, binnen de context waarin coreferentie relevant is, data verspreid over verschillende bronnen best altijd volgens standaarden worden opgeslagen. Veronderstellen we opnieuw een bedrijf met als departementen van dat bedrijf de diensten aankoop, marketing en verkoop. Wanneer deze diensten, elk voor hun specifieke doeleinden, producten beschrijven, dan moet dit op een gestandaardiseerde manier gebeuren, teneinde de complexiteit van coreferentiedetectie over de verschillende databronnen te kunnen minimaliseren. Het gebrek aan standardisatie vormt bijvoorbeeld een belangrijke uitdaging bij ETL-processen in de context van *data warehouses*.

2.3 Onzekerheid bij coreferentie

In dit deel worden een aantal veronderstellingen gemaakt over de context waarin coreferentie in deze thesis verder wordt behandeld. In het algemeen wordt een onderscheid gemaakt tussen twee bronnen van onzekerheid over coreferentie: onzekerheid over structurele aspecten en onzekerheid over inhoudelijke aspecten.

Onzekerheid kan in eerste instantie ontstaan doordat structurele verschillen bestaan in de manier waarop objecten entiteiten gaan beschrijven. Bij complexe objecten wil dit zeggen dat het universum $U_1 \times ... \times U_n$ verschilt voor objecten die men wil vergelijken. Dit probleem kan zich op verschillende

manieren manifesteren[2].

- Wanneer overheen verschillende databanken dezelfde entiteitstypes met een verschillend label worden aangeduid, kunnen de universa waarin entiteiten van types worden beschreven niet meer eenvoudig aan elkaar worden gekoppeld. Anderzijds geeft eenzelfde label overheen verschillende databanken geen garantie dat dezelfde entiteitstypes worden bedoeld.

- De lijst van eigenschappen waarmee entiteiten worden beschreven kan variëren overheen verschillende databanken.

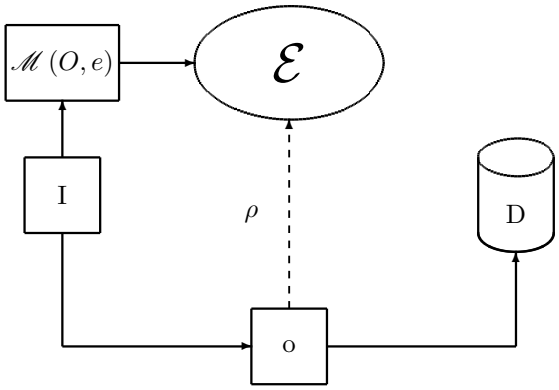
- Gegevens die hetzelfde voorstellen kunnen verschillen van elkaar door arithmetische of functionele transformaties. Overheen complexe universa kunnen geldbedragen bijvoorbeeld in verschillende valuta worden bijgehouden.

Dergelijke situaties creëren elk op hun beurt onzekerheid bij het zoeken naar coreferente objecten. De wetenschappelijke literatuur biedt een aantal oplossingen voor dit probleem [34]. Er kan ruwweg een onderscheid worden gemaakt tussen heuristische technieken en formele technieken. De heuristische technieken worden vooral ingezet om atomaire universa aan elkaar te koppelen. Formele technieken pakken eerder het probleem van datatransformaties aan. Interessante technieken hieromtrent zijn [35, 36]. Het probleem van structurele verschillen in de algemeenheid aanpakken op een automatische wijze is naar onze mening bijzonder moeilijk, in die zin dat een niet te overziene veelheid aan problemen zich kunnen voordoen. Dit wil uiteraard niet zeggen dat het maken van een aantal veronderstellingen niet kan leiden tot een goede benaderende oplossing. Een kernvraag is dan of de gemaakte veronderstellingen niet leiden tot zeer probleemspecifieke oplossingen. Structurele verschillen zullen verder niet worden beschouwd in deze thesis, aangezien ze een totaal ander onderzoeksdomein beslaan dan de inhoudelijke verschillen die hier wel worden bestudeerd. Wanneer we twee objecten vergelijken zal steeds worden verondersteld dat ze deel uitmaken van hetzelfde (complex) universum, dat ze dezelfde entiteitstypes beschrijven en dat ze gelijke labels hebben.

In tweede instantie kan onzekerheid voortkomen uit inhoudelijke verschillen tussen objecten. Om deze oorzaken beter in kaart te brengen, beschouwen we Figuur 2.2 en meer bepaald het meetproces $\mathscr{M}$, dat door het invoerproces I wordt gebruikt om een object o te construeren. Voor een gegeven entiteit e bestaat de constructie van een object o erin om n vragen te beantwoorden van het type 'Wat is de i^{de} eigenschap van e?'. Anders gezegd, de i^{de} eigenschap van entiteit e moet worden gemeten. Formeel gezien kunnen we stellen dat, voor elke entiteit e, het invoerproces I een object $o \in O = U_1 \times ... \times U_n$ construeert zodat:

$$o = \mathscr{M}(O, e) = (\mathscr{M}_1(U_1, e), ..., \mathscr{M}_n(U_n, e)). \tag{2.14}$$

[2]Voor een volledig overzicht van structurele verschillen bij (heterogene) databanken verwijzen we naar [33].



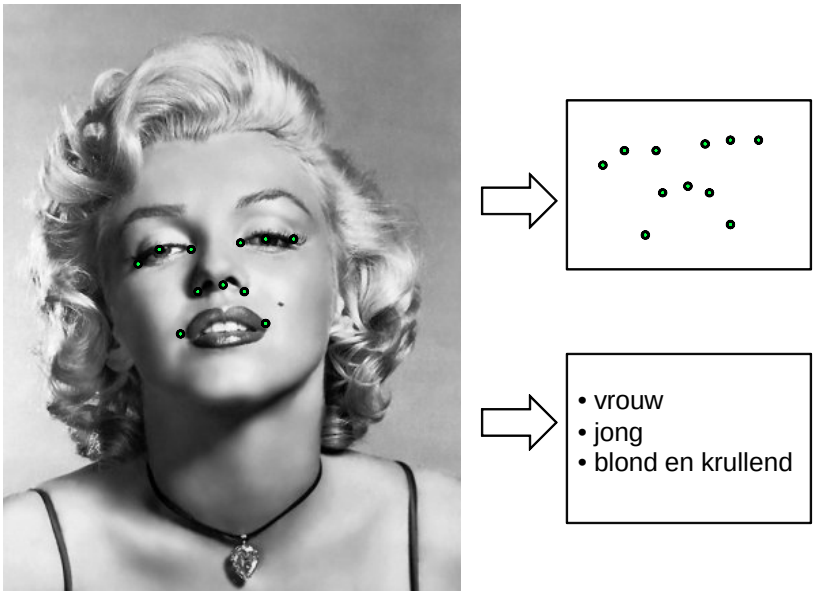
Figuur 2.2: Invoer van data in een databron aan de hand van meetprocessen

Hierbij betekent $\mathscr{M}(O,e)$: 'meet entiteit e en druk deze meting uit in O' en betekent $\mathscr{M}_i(U_i,e)$: 'meet de i^{de} eigenschap van entiteit e en druk deze meting uit in U_i'. Het meten van entiteiten kan enerzijds gebeuren door een entiteit rechtstreeks te meten (bijvoorbeeld het opmaken van een persoonsbeschrijving), maar kan anderzijds ook door transformatie van een object naar een andere ruimte. Stel bijvoorbeeld dat een persoon moet worden beschreven en dat een foto van die persoon beschikbaar is. Deze foto kan worden gezien als een object en kan worden voorgesteld als een collectie van pixels. Deze voorstelling is echter niet bruikbaar voor de vergelijking van personen, waardoor een transformatie naar een andere ruimte nodig is. Hierbij kunnen bijvoorbeeld technieken worden gebruikt voor de extractie van data en/of meta-data uit beelden [37]. Figuur 2.3 toont hoe een dergelijke transformatie werkt en illustreert dat metingen op verschillende manieren kunnen gebeuren.

Onzekerheid over de beslissing dat twee objecten coreferent zijn, wordt veroorzaakt door problemen die zich stellen aangaande de meetprocessen $\mathscr{M}_i$. Dit leidt ons naar een classificatie van onzekerheidsoorzaken.

Een eerste probleem stelt zich wanneer de i^{de} eigenschap niet meetbaar is voor e. Dit kan, althans binnen het hier gegeven kader van coreferentie, twee oorzaken hebben: de eigenschap is onbekend of ze bestaat niet, d.i. ze is niet van toepassing voor entiteit e. We merken op dat een extra moeilijkheid wordt veroorzaakt door het feit dat informatiesystemen in de praktijk meestal geen onderscheid maken tussen deze interpretaties en een niet-meetbare waarde voorstellen door een standaardsymbool. Wanneer niet-meetbare data voorkomen, is het duidelijk dat hieraan aandacht besteed moet worden binnen het kader van coreferentie. Gezien de oorzaak van niet-meetbaarheid in de praktijk niet bekend is, zal er over de oorzaak meestal op voorhand een veronderstelling worden gemaakt.

Een tweede probleem doet zich voor wanneer een eigenschap meetbaar is,



Figuur 2.3: Meting door transformatie

maar het meetproces onnauwkeurig is. Dit probleem stelt zich voornamelijk, hoewel niet exclusief, bij elektrische of mechanische meetprocessen, die haast altijd aan onnauwkeurigheden onderhevig zijn. Echter, voor veel meetprocessen is informatie voorhanden betreffende de meetfout en deze informatie kan in rekening worden gebracht om de onzekerheid te modelleren die ontstaat door de meetfout.

Een derde probleem treedt op wanneer het resultaat van een meting niet eenduidig bepaald is. Wanneer verschillende correcte antwoorden bestaan, is het mogelijk dat in verschillende bronnen een verschillend antwoord wordt gebruikt bij de beschrijving van dezelfde entiteit, hetgeen per definitie onzekerheid veroorzaakt. Het bestaan van dit probleem kan op zich verschillende oorzaken hebben. Ten eerste is de vraag die achter het meetproces schuilgaat niet altijd een objectieve vraag. Wanneer naar een subjectief oordeel wordt gepeild, spreekt het voor zich dat meerdere antwoorden mogelijk zijn. Dit is bijvoorbeeld het geval wanneer er wordt gevraagd wat het type of de prijscategorie van een restaurant is. Ten tweede is het zo dat in sommige gevallen antwoorden verschillen in hun informatiewaarde. Een typisch voorbeeld hiervan doet zich voor bij geografische gebieden, waar een hiërarchie van deelgemeenten, gemeenten en steden kan worden gebruikt. Hierdoor is het beschrijven van een gebied niet altijd eenduidig. Ten derde kan een equivalentierelatie bestaan op U_i met betrekking tot metingen door $\mathscr{M}_i$. De synoniemenrelatie is hiervan een voorbeeld. Een vierde reden kan zijn dat er eenvoudig weg geen consensus bestaat in de opmeting. Dit probleem verschilt van het tweede probleem in die

zin dat het antwoord steeds correct is, daar waar dit voor het tweede probleem niet het geval is. Bij onnauwkeurigheid is er één correct antwoord, maar is het niet mogelijk om dit antwoord te vinden. Beide problemen zijn orthogonaal aan elkaar.

Een vierde probleem dat in rekening moet worden gebracht, is de aanwezigheid van ruis. We houden er hierbij expliciet rekening mee dat ruis kan optreden op verschillende niveau's. Enerzijds kan het meetproces $\mathscr{M}_i$ een ruizig antwoord geven op een vraag $\mathscr{M}_i(U_i, e)$, in die zin dat $\mathscr{M}_i$ de intentie heeft om correct te antwoorden, maar door ruis een fout antwoord geeft. Een typisch voorbeeld hiervan zijn de schrijffouten in karakterstrings, waar de intentie om een bepaald antwoord te geven aanwezig is, maar een schrijffout belet dat het antwoord volledig correct is. Anderzijds kan $\mathscr{M}_i$ op zich ruizig zijn in de productie van antwoorden, waardoor een fout antwoord gegeven wordt op een vraag. In die context heeft $\mathscr{M}_i$ geen intentie om de vraag correct te beantwoorden en resulteert $\mathscr{M}_i$ bewust in een foute meting. Wanneer de nationaliteit van de Belgische tennisspeelster Justine Henin wordt gemeten, dan is het antwoord 'Brits' het resultaat van een foute meting van de eigenschap 'nationaliteit'. Dergelijke ruizige meetprocessen kunnen leiden tot *tegenspraken* in databanken.

Een vijfde aspect van $\mathscr{M}_i$ dat relevant is in de context van coreferentie, is het feit dat $\mathscr{M}_i(U_i, e)$ niet altijd constant is over de tijd. Anders gezegd kan het zijn dat voor twee tijdstippen t en t' $\mathscr{M}_i^t(U_i, e) \neq \mathscr{M}_i^{t'}(U_i, e)$. Enerzijds kan men inzien dat dit probleem in feite wordt gelimiteerd door een goede keuze van de eigenschappen die men gebruikt voor de beschrijving van entiteiten. Zo is de lichaamstemperatuur van een persoon niet geschikt als beschrijvende eigenschap, precies omdat deze eigenschap sterk variabel is doorheen de tijd. Anderzijds, is het perfect mogelijk dat sommige beschrijvende eigenschappen wijzigen doorheen de tijd, hoewel niet frequent. Dit fenomeen staat in de context van *data warehouses* bekend als *traag veranderende dimensies* [38], hetgeen in de context van onze probleembeschrijving kan worden hernoemd tot *traag veranderende eigenschappen*. Een typisch voorbeeld van dergelijke eigenschappen zijn adresgegevens.

Als zesde en laatste probleem kan het zijn dat er onzekerheid in de opslag van data wordt toegestaan. Meer bepaald, indien een eigenschap niet eenvoudig meetbaar is, wordt in possibilistische databanken [39] toegestaan dat een possibiliteitsverdeling π_{U_i} wordt opgeslagen[3]. Dit laat toe om de intrinsieke onzekerheid, gekoppeld aan het antwoord op de vraag $\mathscr{M}_i(U_i, e)$ te modelleren. Het mag duidelijk zijn dat dit probleem gelinkt is aan het tweede probleem dat we hierboven beschrijven, namelijk dat van onnauwkeurige meetprocessen. Inderdaad kunnen we dit tweede probleem in een ruimere context beschouwen als een vraag waarop het antwoord onzeker is omwille van een onnauwkeurige meting. In de nauwere context wordt echter maar één antwoord opgeslagen, hetgeen een gevolg is van beperkingen in de opslagfaciliteiten van de databank. Dit is in de ruimere context van possibilistische databanken anders. In een possibilistische databank wordt voor elke $u_i \in U_i$ aangegeven

[3] Als alternatief kunnen ook probabilistische databanken worden beschouwd.

wat de mogelijkheid is dat u_i een correct antwoord is op de gestelde vraag. Aan het probleem van onzekerheid in opslag van data zal minder aandacht worden besteed om verschillende redenen. Ten eerste staat de implementatie van commerciële possibilistische databanken nog in de kinderschoenen in vergelijking met bestaande systemen voor dataopslag. Rekening houdend met de moeite die objectgeoriënteerde databanksystemen hebben om een ruim marktaandeel in te winnen, is het zeer de vraag of possibilistische databanken ooit een marktdoorbraak zullen kennen. Dit mag op zich natuurlijk geen reden zijn om het onderzoek ernaar aan de kant te schuiven. Er speelt echter een bijkomende reden. Het vergelijken van possibiliteitsverdelingen met als antwoord een possibilistische waarheidswaarde, is reeds uitgebreid bestudeerd [8]. We zullen aantonen dat deze oplossingen volledig compatibel zijn met het raamwerk voor coreferentie dat hier wordt geconstrueerd. Met andere woorden, we zullen aantonen dat de bestaande operatoren gebruikt kunnen worden in de context van coreferentie.

De classificatie van imperfecties die kunnen optreden bij $\mathscr{M}_i$ zal verderop worden gebruikt ter verificatie van de voorgestelde oplossingen. De vermelde classificatie steunt op de classificatie van Motro [40, 41].

2.4 Evaluatoren

De gegeven problematiek van coreferentie zal in deze thesis vanuit een possibilistisch standpunt worden benaderd. Andere werken in de literatuur benaderen het probleem vanuit een probabilistisch standpunt [42]. Laat ons daarom eerst een motivatie geven voor het gebruik van possibiliteiten. Gelet op de definitie van objectreferentie (Definitie 2.2) en coreferentie (Definitie 2.10), komen we tot de conclusie dat het bepalen van coreferentie kan worden gezien als een Boolese vraag waarop een antwoord moet worden geformuleerd. Anders gezegd, gegeven twee objecten, dan is de vraag 'Zijn deze twee objecten coreferent?' zinvol en het antwoord is 'Ja' of 'Neen'. Voor diezelfde objecten is de vraag 'Wat is de mate van coreferentie van deze twee objecten?' niet zinvol. Het antwoord op de Boolese vraag kan echter om verschillende redenen onzeker zijn, waardoor bepaling van coreferentie niet triviaal is. Gelet op het onderscheid dat Dubois en Prade maken tussen types onzekerheid (Hoofdstuk 1), behoort de onzekerheid omtrent coreferentie tot de categorie van onzekerheid door onvolledigheid. Deze observatie volgt ook uit de schematische voorstelling van coreferentie in Figuur 2.2, waaruit blijkt dat coreferentiedetectie neerkomt op de reconstructie van de onbekende functie ρ. Hieruit volgt dat het zinvol is om een functie te beschouwen die, gegeven twee objecten, de onzekerheid schetst over het antwoord op de vraag of deze twee objecten coreferent zijn en die deze onzekerheid op een possibilistische manier voorstelt.

Definitie 2.11 (Evaluator)
Gegeven een universum van objecten O, dan bestaat er voor elk koppel van waarden $(o_1, o_2) \in O^2$ een bevestigende propositie p:

$$p_{(o_1,o_2)} = \text{“}o_1 \text{ en } o_2 \text{ zijn coreferent”}$$

De propositie $p_{(o_1,o_2)}$ wordt de coreferentiële propositie van o_1 en o_2 genoemd. Een evaluator over O wordt genoteerd als E_O en is gedefinieerd als:

$$\begin{array}{rcl} E_O : O^2 & \rightarrow & \mathcal{F}(\mathbb{B}) : \\ (o_1, o_2) & \mapsto & E_O(o_1, o_2) = \left\{ \left(T, \mu_{\widetilde{p}_{(o_1,o_2)}}(T) \right), \left(F, \mu_{\widetilde{p}_{(o_1,o_2)}}(F) \right) \right\} \end{array} \tag{2.15}$$

waarbij $\widetilde{p}_{(o_1,o_2)}$ de onzekerheid over de waarheidswaarde van $p_{(o_1,o_2)}$ voorstelt, d.i. $\mu_{\widetilde{p}_{(o_1,o_2)}}(T)$ geeft de mogelijkheid dat $p_{(o_1,o_2)}$ waar is en $\mu_{\widetilde{p}_{(o_1,o_2)}}(F)$ geeft de mogelijkheid dat $p_{(o_1,o_2)}$ vals is.

Een evaluator is het fundamentele concept waarrond deze thesis is opgebouwd. In de rest van dit hoofdstuk worden evaluatoren eerst in de algemeenheid bestudeerd. In het bijzonder wordt een vergelijking gemaakt tussen de eigenschappen die van een evaluator verondersteld mogen worden en de eigenschappen die intrisiek zijn aan de relatie $\leftrightarrow$. Ook worden in dit hoofdstuk enkele aspecten van evaluatoren bestudeerd die los staan van de gebruikte objectruimte.

Definitie 2.11 geeft duidelijk een zeer algemene omschrijving van wat een evaluator precies is. Meer bepaald worden geen beperkingen opgelegd aan de uitgedrukte onzekerheid in functie van de objectwaarden die worden vergeleken. Laat ons daarom beginnen met te onderzoeken of een aantal bijkomende eigenschappen voor evaluatoren kunnen worden toegevoegd aan de algemene definitie. We zullen aangeven in welke situaties deze eigenschappen verondersteld mogen worden. Het zal daaruit blijken dat de vermelde eigenschappen niet beschouwd kunnen worden in Definitie 2.11 zonder de algemeenheid te schaden. Een goed vertrekpunt voor deze studie, zijn de eigenschappen waaraan de coreferentierelatie $\leftrightarrow$ (een equivalentierelatie over O) moet voldoen: reflexiviteit, symmetrie en transitiviteit.

Laat ons beginnen met reflexiviteit te bestuderen. Een eerste belangrijke vraag die zich stelt is, waarom deze eigenschap niet axiomatisch kan worden opgelegd aan evaluatoren. De reden hiervoor is beschreven in de tweede regel voor goed ontwerp. Wanneer een ontoereikende granulariteit wordt beschouwd voor de beschrijving van entiteiten, kan het zijn dat gelijke objecten niet naar eenzelfde entiteit refereren. In dit geval is het dus niet houdbaar om reflexiviteit te veronderstellen. De regel voor goed ontwerp omtrent granulariteit stelt dat objecten een beschrijving van entiteiten moeten zijn op die manier dat de mogelijkheid voor niet-coreferentie van gelijke objecten verwaarloosbaar klein wordt. Wanneer aan deze voorwaarde is voldaan, kan worden gesteld dat een evaluator reflexief moet zijn.

Een tweede reden om reflexiviteit niet in de definitie van een evaluator op te nemen, is het feit dat beschrijvende data in sommige gevallen een equivalentieklasse beschrijven in $\mathcal{E}$, eerder dan een element van $\mathcal{E}$. Dit probleem stelt zich bijvoorbeeld als $\mathcal{E}$ een niet-aftelbare verzameling is, waarbij een ononderscheidbaarheid[4] over elementen van $\mathcal{E}$ met betrekking tot de voorstellingswijze in de

[4]Onnauwkeurigheid wordt veroorzaakt door meetfouten, daar waar ononderscheidbaarheid te wijten is aan de grenzen van het meetproces.

objectwereld aanwezig is. Denken we hierbij bijvoorbeeld aan de reële getallen of de verzameling van kleuren. In zulke gevallen kan $\mathcal{E}$ steeds aftelbaar gemaakt worden door een partitie op $\mathcal{E}$ te beschouwen. Echter, vanuit de strikte definitie die aan objectreferentie gegeven wordt, is het perfect mogelijk dat een object twee verschillende, doch equivalente entiteiten beschrijft. Het is bijvoorbeeld mogelijk dat twee verschillende kleuren eenzelfde RGB code hebben. Reflexiviteit kan dan wederom niet worden verondersteld. Echter, in dit laatste geval is men vaak niet geïnteresseerd in welk element exact wordt gerefereerd, maar wel in de equivalentieklasse die wordt gerefereerd. Dit omdat verschillende entiteiten per definitie niet te onderscheiden zijn. In dergelijke situaties zullen we steeds impliciet een gepartitioneerde entiteitenruimte veronderstellen. In die zin kunnen we dan een evaluator over de gepartitioneerde ruimte beschouwen, waarbij deze evaluator reflexief mag worden verondersteld.

Merken we op dat er zich in beide gevallen een equivalentierelatie vormt over de entiteitenruimte, zij het om verschillende redenen. Een ontoereikende granulariteit creëert een equivalentierelatie over $\mathcal{E}$ door het verlies aan informatie, daar waar bij ononderscheidbaarheid ten opzichte van de voorstellingswijze deze equivalentierelatie impliciet aanwezig is. Tenzij één van voornoemde gevallen zich voordoet, kunnen we een reflexieve evaluator veronderstellen, die als volgt wordt gedefinieerd.

Definitie 2.12 (Reflexieve evaluator)
Gegeven een universum O. Een reflexieve evaluator E_O is een evaluator over O waarvoor geldt:

$$\forall(o_1, o_2) \in O^2 : o_1 = o_2 \Rightarrow E_O(o_1, o_2) = (1, 0). \tag{2.16}$$

Hierbij is $(1, 0)$ een possibilistische waarheidswaarde die stelt dat p zeker waar is.

Onder de veronderstelling van een reflexieve evaluator kan ρ als functie worden gereconstrueerd. Gelijke objecten worden door een reflexieve evaluator altijd beschouwd als zijnde zeker coreferente objecten. Omgekeerd is het onder de beperking van reflexiviteit echter niet zo dat objecten waarvoor coreferentie een zekerheid is, ook gelijke objecten zijn. Deze idee wordt gedreven door eventuele onnauwkeurigheden bij het meetproces, waardoor het mogelijk is om, rekening houdend met meetfouten, te stellen dat twee verschillende objecten, toch coreferent zijn. Bij afwezigheid van dergelijke onnauwkeurigheden, kan het gewenst zijn om een sterkere beperking op te leggen aan een evaluator, met name sterke reflexiviteit.

Definitie 2.13 (Sterk reflexieve evaluator)
Gegeven een universum O. Een sterk reflexieve evaluator E_O is een evaluator over O waarvoor geldt:

$$\forall(o_1, o_2) \in O^2 : o_1 = o_2 \Leftrightarrow E_O(o_1, o_2) = (1, 0). \tag{2.17}$$

Als volgende eigenschap bespreken we symmetrie van de evaluator. Opnieuw beginnen we met een antwoord te geven op de vraag waarom deze eigenschap niet in Definitie 2.11 vervat zit. Beschouwen we twee databronnen A en B met objecten die entiteiten uit een universum $\mathcal{E}$ beschrijven, onder de veronderstelling dat beide databronnen een identiek objectuniversum gebruiken. Dan stellen we vast dat symmetrie van evaluatoren enkel verondersteld mag worden als de processen $\mathscr{M}$ en I voor beide databronnen identiek zijn. Meer bepaald moet het model van imperfecties gelijk zijn. Echter, als $\mathscr{M}(O, e)$ niet objectief kan worden beantwoord, dan wil dit zeggen dat het antwoord op deze vraag vanuit een bepaalde subjectieve context gegeven wordt en dat deze context verschillend kan zijn voor A en B. In dat geval moet er bij de bepaling van onzekerheid rekening worden gehouden met de context waarin een antwoord op de vraag $\mathscr{M}(O, e)$ is gegeven. We kunnen echter inzien dat veel situaties bestaan waarin de processen $\mathscr{M}$ en I identiek zijn. Ten eerste zijn objectieve antwoorden niet aan deze problematiek onderhevig en ten tweede is het niet onredelijk om de subjectieve contexten als identiek te veronderstellen. Onder deze veronderstellingen kunnen we een symmetrische evaluator definiëren.

Definitie 2.14 (Symmetrische evaluator)
Gegeven een universum O. Een symmetrische evaluator E_O is een evaluator over O waarvoor geldt:

$$\forall (o_1, o_2) \in O^2 : E_O(o_1, o_2) = E_O(o_2, o_1). \tag{2.18}$$

Tenzij anders vermeld, zullen we steeds veronderstellen dat een evaluator symmetrisch is.

Als laatste eigenschap bestuderen we hoe de transitiviteit van de relatie $\leftrightarrow$ kan worden overdragen op evaluatoren. Gegeven drie objecten o_1, o_2 en o_3. De vereiste van transitiviteit van $\leftrightarrow$ impliceert dat er geldt dat:

$$(o_1 \leftrightarrow o_2 \wedge o_2 \leftrightarrow o_3) \leq (o_1 \leftrightarrow o_3). \tag{2.19}$$

Merk op dat het $\leq$-teken volgt uit de natuurlijke ordening op $\mathbb{B}$ gecombineerd met de waarheidstabel van een logische implicatie. Aangezien dit moet gelden voor elk willekeurig triple van elementen, kunnen we de volgende implicaties afleiden:

$$\begin{array}{rcrcr} (o_1 \leftrightarrow o_2) & \wedge & (o_2 \leftrightarrow o_3) & \Rightarrow & (o_1 \leftrightarrow o_3) \\ (o_1 \leftrightarrow o_2) & \wedge & \neg(o_2 \leftrightarrow o_3) & \Rightarrow & \neg(o_1 \leftrightarrow o_3) \\ \neg(o_1 \leftrightarrow o_2) & \wedge & (o_2 \leftrightarrow o_3) & \Rightarrow & \neg(o_1 \leftrightarrow o_3)\,. \end{array} \tag{2.20}$$

Laat ons (2.20) verder onderzoeken wanneer onzekerheid in beschouwing wordt genomen. Veronderstellen we daarvoor de possibilistische waarheidswaarden $\widetilde{p}_{(o_1,o_2)}$ en $\widetilde{p}_{(o_2,o_3)}$ die onzekerheid uitdrukken over de coreferentie van de koppels (o_1, o_2) en (o_2, o_3). Hierbij veronderstellen we dat de onzekerheid geproduceerd is door eenzelfde evaluator E_O. Beschouwen we nu de onzekerheid $\widetilde{p}_{(o_1,o_3)}$. Enerzijds kunnen we, vertrekkende van een relatie g, het uitbreidingsprincipe van Zadeh toepassen, zodat we een possibilistische vertaling $\widetilde{g}$ verkrijgen. Deze relatie:

$$g : \mathbb{B}^2 \to \mathcal{P}(\mathbb{B}) \tag{2.21}$$

wordt gegeven door:

$$\begin{aligned} g(T,T) &= \{T\} \\ g(T,F) &= \{F\} \\ g(F,T) &= \{F\} \\ g(F,F) &= \{T,F\}. \end{aligned} \tag{2.22}$$

Merk op dat g géén functie is aangezien het koppel (F,F) verschillende beelden heeft onder g. Dit strookt met het feit dat, als zowel (o_1,o_2) en (o_2,o_3) geen coreferente koppels zijn, er geen informatie voorhanden is over de coreferentie van o_1 en o_3. Het gebruik van een relatie eerder dan een functie leidt ons dus reeds naar een binaire possibilistische aanpak. Passen we op g het uitbreidingsprincipe van Zadeh toe, dan krijgen we de Zadeh-uitbreiding $\widetilde{g}$ van g:

$$\widetilde{g} : \mathcal{F}(\mathbb{B})^2 \to \mathcal{F}(\mathbb{B}) \tag{2.23}$$

waarvoor, voor willekeurige $\widetilde{p}$ en $\widetilde{q}$, geldt dat:

$$\mu_{\widetilde{g}(\widetilde{p},\widetilde{q})}(T) = \max \begin{pmatrix} \min\left(\mu_{\widetilde{p}}(T), \mu_{\widetilde{q}}(T)\right), \\ \min\left(\mu_{\widetilde{p}}(F), \mu_{\widetilde{q}}(F)\right) \end{pmatrix} \tag{2.24}$$

$$\mu_{\widetilde{g}(\widetilde{p},\widetilde{q})}(F) = \max \begin{pmatrix} \min\left(\mu_{\widetilde{p}}(T), \mu_{\widetilde{q}}(F)\right), \\ \min\left(\mu_{\widetilde{p}}(F), \mu_{\widetilde{q}}(T)\right), \\ \min\left(\mu_{\widetilde{p}}(F), \mu_{\widetilde{q}}(F)\right) \end{pmatrix}. \tag{2.25}$$

De Zadeh-uitbreiding $\widetilde{g}$ van g geeft ons een grens voor de onzekerheid die aanwezig is over de coreferentie van o_1 en o_3. Met andere woorden:

$$\forall b \in \mathbb{B} : \mu_{E_O(o_1,o_3)}(b) \leq \mu_{\widetilde{g}\left(\widetilde{p}_{(o_1,o_2)},\widetilde{p}_{(o_2,o_3)}\right)}(b). \tag{2.26}$$

Een andere manier om aan dit resultaat te komen, is te redeneren in termen van necessiteiten. Gebruiken we de volgende notaties:

$$\mathcal{N}(p_{(o_1,o_3)} = T) = \mathrm{Nec}(p_{(o_1,o_3)} = T | \widetilde{p}_{(o_1,o_2)}, \widetilde{p}_{(o_2,o_3)}) \tag{2.27}$$

$$\mathcal{N}(p_{(o_1,o_3)} = F) = \mathrm{Nec}(p_{(o_1,o_3)} = F | \widetilde{p}_{(o_1,o_2)}, \widetilde{p}_{(o_2,o_3)}) \tag{2.28}$$

waarbij $\mathrm{Nec}(A|C)$ de conditionele necessiteit van A gegeven C voorstelt (Hoofdstuk 1). Steunend op wat we weten uit (2.20) krijgen we:

$$\mathcal{N}(p_{(o_1,o_3)} = T) \geq \min\left(\mathrm{Nec}(p_{(o_1,o_2)} = T), \mathrm{Nec}(p_{(o_2,o_3)} = T)\right) \tag{2.29}$$

$$\mathcal{N}(p_{(o_1,o_3)} = F) \geq \min\left(\mathrm{Nec}(p_{(o_1,o_2)} = T), \mathrm{Nec}(p_{(o_2,o_3)} = F)\right) \tag{2.30}$$

$$\mathcal{N}(p_{(o_1,o_3)} = F) \geq \min\left(\mathrm{Nec}(p_{(o_1,o_2)} = F), \mathrm{Nec}(p_{(o_2,o_3)} = T)\right). \tag{2.31}$$

Het kan door een gevallenstudie worden aangetoond dat deze uitdrukkingen in overeenstemming zijn met (2.26). In de praktijk is de constructie van een evaluator die hieraan voldoet, niet eenvoudig. Als we een automatische oplossing willen opstellen voor het vinden van coreferente objecten, moeten we

veronderstellen dat dit koppelsgewijs mogelijk is, vertrekkende van een verzameling van eenvoudige kennisregels. In de praktijk vormen deze kennisregels een benadering van de imperfecties van een meetproces $\mathscr{M}$. In vele praktische gevallen zal blijken dat deze kennisregels daarom op zich onvolledige kennis voorstellen. De kennis die ter beschikking wordt gesteld om coreferente objecten te identificeren, volstaat niet in een dergelijk geval. Hierdoor kan een evaluator inconsistente beslissingen nemen. In die zin is (2.26) een hulpmiddel om dergelijke inconsistenties op te sporen en zullen we haar niet als een axiomatische vereiste behandelen. Voor een verdere bespreking van inconsistenties verwijzen we naar Hoofdstuk 5. Wanneer aan (2.26) voldaan is, spreken we van een transitieve evaluator.

Definitie 2.15 (Transitieve evaluator)
Gegeven een universum O. Een transitieve evaluator E_O is een evaluator over O waarvoor er geldt dat, voor willekeurige o_1, o_2 en o_3:

$$1 - \mu_{E_O(o_1,o_3)}(F) \geq \min\left(1 - \mu_{E_O(o_1,o_2)}(F), 1 - \mu_{E_O(o_2,o_3)}(F)\right) \quad (2.32)$$

$$1 - \mu_{E_O(o_1,o_3)}(T) \geq \min\left(1 - \mu_{E_O(o_1,o_2)}(F), 1 - \mu_{E_O(o_2,o_3)}(T)\right) \quad (2.33)$$

$$1 - \mu_{E_O(o_1,o_3)}(T) \geq \min\left(1 - \mu_{E_O(o_1,o_2)}(T), 1 - \mu_{E_O(o_2,o_3)}(F)\right). \quad (2.34)$$

De gevolgen van de transitiviteit van E_O kunnen aanschouwelijk worden voorgesteld aan de hand van een gevallenstudie. Beschouwen we een transitieve evaluator E_O, $(o_1, o_2, o_3) \in O^3$ en $(a, b, c) \in [0, 1]^3$. De transitiviteit van E_O kan worden samengevat in de volgende drie gevallen.

(1) Als er geldt dat:

$$(E_O(o_1, o_2) = (1, a)) \wedge (E_O(o_2, o_3) = (1, b)) \quad (2.35)$$

dan volgt er uit de transitiviteit van E_O dat:

$$E_O(o_1, o_3) = (1, \max(a, b)). \quad (2.36)$$

(2) Als er geldt dat:

$$(E_O(o_1, o_2) = (1, a)) \wedge (E_O(o_2, o_3) = (b, 1)) \quad (2.37)$$

en bovendien is $a \geq b$, dan volgt er uit de transitiviteit van E_O dat:

$$E_O(o_1, o_3) = (c, 1) \quad (2.38)$$

met als voorwaarde dat $a \geq c$.

(3) Als er geldt dat:

$$(E_O(o_1, o_2) = (1, a)) \wedge (E_O(o_2, o_3) = (b, 1)) \quad (2.39)$$

en bovendien is $a < b$, dan volgt er uit de transitiviteit van E_O dat:

$$E_O(o_1, o_3) = (b, 1). \quad (2.40)$$

Indien aan (2.26) niet voldaan is, kunnen zich twee gevallen voordoen. In een eerste geval voldoet een evaluator niet aan (2.26), doordat de evaluator meer onzekerheid hecht aan de coreferentie van o_1 en o_3, dan kan worden afgeleid via (2.26). De evaluator uit dan meer onzekerheid dan in principe kan worden afgeleid uit de kennis die voorhanden is. In dit geval zal de meest mogelijke waarheidswaarde aangegeven door rechtstreekse vergelijking, overeenstemmen met de meest mogelijke waarheidswaarde die afgeleid kan worden uit (2.26). In een tweede geval kan de meest mogelijke waarheidswaarde aangegeven door E_O verschillen van de meest mogelijke waarheidswaarde aangegeven door (2.26). Het onderscheid tussen deze twee gevallen geeft aanleiding tot de volgende definitie.

Definitie 2.16 (Consistente evaluator)
Gegeven een universum O. Een consistente evaluator E_O is een evaluator over O waarvoor er geldt dat, voor willekeurige o_1, o_2 en o_3, als:

$$\begin{array}{ll} \widetilde{p} & = E_O(o_1, o_3) \\ \widetilde{q} & = \widetilde{g}(E_O(o_1, o_2), E_O(o_2, o_3)) \end{array} \tag{2.41}$$

dan:

$$(\mu_{\widetilde{p}}(T) = 1 \wedge \mu_{\widetilde{q}}(T) = 1) \vee (\mu_{\widetilde{p}}(F) = 1 \wedge \mu_{\widetilde{q}}(F) = 1)\,. \tag{2.42}$$

Stelling 2.1
Gegeven een universum O. Een evaluator E_O over O is consistent als er een partitie van O bestaat, zodat er geldt dat:

$$\forall (o_1, o_2) \in O_i \times O_j \quad : \quad (i \neq j) \Rightarrow (\mu_{E_O(o_1,o_2)}(F) = 1) \tag{2.43}$$

$$\forall (o_1, o_2) \in O_i^2 \quad : \quad \mu_{E_O(o_1,o_2)}(T) = 1 \tag{2.44}$$

Bewijs Triviaal. □

In het algemeen zal een onderscheid worden gemaakt tussen twee klassen van evaluatoren: syntactische evaluatoren en semantische evaluatoren. Syntactische evaluatoren zijn evaluatoren die onzekerheid uitdrukken op basis van een syntactische vergelijking tussen twee objecten. Hiermee wordt bedoeld dat deze evaluatoren een vergelijking maken op basis van syntactische kenmerken van objecten. Zo kan een syntactische evaluator gebruikt worden om bijvoorbeeld de karakterstrings “Lucas” en “Lukas” of de reële waarden 1.232 en 1.233 te vergelijken. Semantische evaluatoren vertrekken van een binaire relatie R over het universum O om hun onzekerheid te formuleren. Zo kunnen ze bijvoorbeeld een vergelijking maken tussen de karakterstrings “New York” en “Manhattan”. Merk op dat het onderscheid tussen deze twee klassen van evaluatoren niet altijd éénduidig is. Syntactische evaluatoren construeren een binaire relatie over O door gebruik te maken van kennisregels om syntactische verbanden vast te leggen. Semantische evaluatoren construeren zelf echter geen binaire relatie over O: ze gebruiken een binaire relatie over O om zo objecten te kunnen vergelijken. We zullen dieper ingaan op semantische evaluatoren verderop in dit hoofdstuk.

Twee speciale evaluatoren die in het vervolg van dit werk regelmatig zullen worden aangehaald, zijn de tweewaardige en de driewaardige evaluator.

Definitie 2.17 (Tweewaardige evaluator)
Een tweewaardige evaluator E_O is een evaluator waarvoor geldt:

$$\forall(o_1, o_2) \in O^2 : o_1 \neq o_2 \Rightarrow E_O(o_1, o_2) = (0, 1). \tag{2.45}$$

Definitie 2.18 (Driewaardige evaluator)
Een driewaardige evaluator E_O is een sterk reflexieve evaluator waarvoor er geldt dat:

$$\forall(o_1, o_2) \in O^2 : E_O(o_1, o_2) \in \{(1, 0), (1, 1), (0, 1)\}. \tag{2.46}$$

Tweewaardige en driewaardige evaluatoren zijn sterk reflexief en symmetrisch. Tweewaardige evaluatoren zijn bovendien transitief en driewaardige evaluatoren zijn consistent.

2.5 Moeilijk-meetbaarheid van eigenschappen

Hernemen we de verschillende oorzaken van onzekerheid, dan stellen we vast dat de meeste van deze problemen contextspecifiek zijn. Echter, het probleem van moeilijk-meetbare eigenschappen onderscheidt zich van de andere oorzaken, doordat het los staat van zowel het beschouwde objectenuniversum, als de eigenschap die wordt gemeten. Het probleem van moeilijk-meetbaarheid kan bijgevolg algemeen worden opgelost. We behandelen eerst het bijzonder geval van niet-meetbaarheid, gezien het belang ervan in vele bestaande systemen voor dataopslag. Daarna behandelen we het algemene geval van moeilijk-meetbaarheid.

2.5.1 Niet-meetbare eigenschappen

De aanpak van dit probleem vertrekt vanuit de vaststelling dat niet-meetbaarheid van een eigenschap van een entiteit verschillende oorzaken kan hebben. In een rapport van een ANSI werkgroep wordt vermelding gemaakt van veertien verschillende oorzaken van niet-meetbaarheid [43]. Later is door Codd een vereenvoudigde classificatie met twee categorieën voorgesteld [44, 45, 46]:

1. Een eigenschap is niet van toepassing voor een gegeven entiteit. Het is in dit geval niet zinvol om de eigenschap te meten.

2. Een eigenschap kan niet worden geobserveerd. Het is zinvol om de eigenschap te meten, maar door een belemmering van het meetproces is het resultaat van de meting onbekend.

Ondanks de verschillende oorzaken van niet-meetbaarheid, behandelen traditionele databanksystemen niet-meetbare data steeds op eenzelfde manier.

Ofwel wordt een *standaardwaarde* uit het universum geselecteerd om niet-gespecificeerde gegevens in te vullen, ofwel gebruikt men een vooraf afgesproken symbool om aan duiden dat een bepaalde eigenschap niet meetbaar is. Dit symbool is traditioneel gegeven door "null" en wordt ook in moderne programmeertalen gebruikt. Wij zullen in het vervolg de kortere notatie $\perp$ voor dit symbool hanteren. Een probleem dat zich stelt, is dat het gebruik van één symbool niet langer toelaat om de precieze interpretatie van $\perp$ te achterhalen. We kunnen uiteraard op voorhand veronderstellen dat $\perp$ een welbepaalde interpretatie heeft binnen een gegeven context, maar dit veroorzaakt een aantal problemen. Eerst en vooral is het zo dat voor verschillende eigenschappen van een entiteit een andere interpretatie van niet-meetbaarheid kan gelden. Een tweede en ernstiger probleem is dat dezelfde eigenschap voor verschillende entiteiten niet meetbaar kan zijn omwille van verschillende redenen. De eigenschap 'zwanger' is voor de ene persoon niet van toepassing omdat het een man is, maar kan onbekend zijn voor een vrouw. Het is duidelijk dat een a priori keuze van de interpretatie van $\perp$ niet toelaat om dit tweede probleem op te lossen. Codd heeft als oplossing voor dit probleem voorgesteld om een verschillend symbool voor elke interpretatie van niet-meetbaarheid te gebruiken.

Los hiervan kan een methode worden opgebouwd die aangeeft hoe een evaluator zich moet gedragen wanneer de waarde $\perp$ optreedt. Een methode hiervoor is gegeven in [47], in de context van databankbevraging. Een dergelijke methode kan perfect worden overgedragen naar evaluatoren zoals hier ingevoerd. We veronderstellen een complex universum O waarbij deelobjecten van een object o eigenschappen van een entiteit e beschrijven. Stel U een willekeurig deeluniversum van O en veronderstel een evaluator E_U over U. Een object uit U beschrijft steeds een welbepaalde eigenschap van een entiteit e. We bestuderen hier wat de impact is op de vergelijking van twee objecten uit U als de gemeten eigenschap niet meetbaar is voor sommige entiteiten. Laten we eerst het geval bekijken waar slechts één van de vergeleken objecten gelijk is aan $\perp$. Het andere object veronderstellen we een willekeurig object uit het universum U te zijn. Stel dat we het bestaande object vergelijken met $\perp$, meer bepaald onder de interpretatie van onbekendheid, genoteerd[5] als $\perp_\forall$. Beide objecten beschrijven dan een bestaande eigenschap en het is perfect mogelijk dat de objecten coreferent zijn, maar het is eveneens perfect mogelijk dat de objecten niet coreferent zijn. De evaluator kan in dit geval geen uitspraak doen. Vergelijken we het bestaande object met $\perp$ onder de interpretatie van onbestaand, genoteerd[6] als $\perp_\nexists$, dan stellen we vast dat het resultaat van de evaluator is dat deze objecten onmogelijk coreferent kunnen zijn. Immers, het ene object beschrijft een eigenschap die van toepassing is, het andere object beschrijft een eigenschap die niet van toepassing is. Tabel 2.1 vat deze bevindingen samen.

Vergelijken we de uitkomsten van twee metingen van niet-meetbare eigenschappen, dan kunnen er zich drie scenario's voordoen. Vergelijking van $\perp_\forall$

[5] De notatie $\forall$ komt voort uit het possibilistische principe dat 'onbekend' betekent dat elke waarde uit het universum een mogelijk antwoord op de meting is.

[6] De notatie $\nexists$ komt voort uit het principe dat 'onbestaand' betekent dat geen enkele waarde uit het universum een mogelijk antwoord op de meting is.

met zichzelf verschilt enkel van vergelijking met een bestaand object in de zin dat nog meer onzekerheid wordt toegevoegd. Het resultaat is volstrekt onzeker. Vergelijking van $\perp_\forall$ met $\perp_\nexists$ laat ons besluiten dat de objecten niet coreferent kunnen zijn, wegens de tegenspraak van een zinvolle meting versus niet-zinvolle meting. Vergelijken we $\perp_\nexists$ met zichzelf, dan vergelijken we twee objecten die elk een eigenschap van een entiteit beschrijven, waarbij de eigenschap niet van toepassing is. Het resultaat is dat deze objecten noodzakelijk coreferent moeten zijn, onder de veronderstelling van reflexieve evaluatie. Ook deze bevindingen worden samengevat in Tabel 2.1. We merken op dat deze bevindingen ook formeel kunnen worden afgeleid mits de correcte interpretaties van onbekend en onbestaand in een possibilistisch raamwerk en mits de toepassing van het uitbreidingsprincipe van Zadeh op de gelijkheidsrelatie in O. Deze afleidingen zijn het onderwerp van de volgende sectie.

	$u_1 \in U$	$u_1 = \perp_\forall$	$u_1 = \perp_\nexists$
$u_2 \in U$	$\widetilde{p}$	(1,1)	(0,1)
$u_2 = \perp_\forall$	(1,1)	(1,1)	(0,1)
$u_2 = \perp_\nexists$	(0,1)	(0,1)	(1,0)

Tabel 2.1: Evaluatie in het geval van niet-meetbare eigenschappen

2.5.2 Moeilijk-meetbare eigenschappen

Tussen het geval van perfect meetbare eigenschappen en het geval van niet-meetbare eigenschappen, ligt het geval van moeilijk-meetbare eigenschappen. Zoals eerder aangehaald heeft het ontstaan van possibilistische en probabilistische databanken de weg geopend naar databronnen waarin het modelleren van intrinsieke onzekerheid in opgeslagen waarden mogelijk wordt gemaakt. Wanneer de principes van de possibiliteitstheorie worden toegepast, kan een eigenschap als moeilijk meetbaar worden behandeld en wordt het antwoord van een meting opgeslagen als een possibiliteitsverdeling over het beschouwde universum. We merken op dat meetbare en niet-meetbare eigenschappen in dit kader beiden een speciaal geval zijn. Een belangrijke vraag is nu: wat verstaan we onder moeilijk-meetbare eigenschappen? Hier interpreteren we dit als eigenschappen waarvoor het resultaat van de meting niet duidelijk is. Bij een moeilijk-meetbare eigenschap bestaan er binnen het universum waarin de meting wordt uitgedrukt een aantal objecten die mogelijks het resultaat zijn van de meting. Belangrijk hierbij is dat slechts één resultaat het juiste is, maar door een gebrek aan kennis kan dit resultaat niet worden gegeven. Laat ons dezelfde veronderstellingen als bij niet-meetbaarheid maken, maar met het deeluniversum aangeduid als U_i. We weten nu dat $\mathscr{M}_i(U_i, e) \in U_i$. Echter, het precieze resultaat van de meting is niet gekend. We veronderstellen daarom een possibilistische variabele X_e die waarden aanneemt in U_i en we veronderstellen dat er een possibiliteitsverdeling π_{X_e} gegeven is zodat $\pi_{X_e}(u_i)$ de mogelijkheid voorstelt dat u_i het correcte resultaat van $\mathscr{M}_i(U_i, e)$ is.

Laat ons nu een evaluator E_O veronderstellen die objecten uit O kan vergelijken en laat ons twee possibilistische variabelen X_1 en X_2 veronderstellen die waarden aannemen in O. Wanneer we veronderstellen dat deze variabelen t-onafhankelijk zijn, dan wordt de mogelijkheid dat de combinatie van deze variabelen een bepaald koppel van waarden aanneemt, bepaald door (1.73). Bovendien kan evaluator E_O voor elk koppel $(o_1, o_2) \in O^2$ een uitdrukking geven aan de mogelijkheid dat deze objecten (niet) coreferent zijn. Voor elk koppel van objecten uit O, wordt op deze manier een marginale en een conditionele possibiliteit verkregen die we kunnen combineren door te steunen op (1.68). Merken we op dat de verzameling van mogelijke koppels dient gezien te worden als een disjunctieve verzameling in die zin dat slechts één koppel correct is. In het kader van de possibiliteitstheorie moeten we het koppel kiezen waarvoor de possibiliteit wordt gemaximaliseerd. Dit leidt tot de volgende definitie.

Definitie 2.19 (Evaluator voor possibilistische variabelen)
Gegeven een willekeurig objectuniversum O en $\mathcal{X}$ het universum van possibilistische variabelen die waarden aannemen in O. Een evaluator $E_{\mathcal{X}}$ wordt geïnduceerd door een evaluator E_O over O waarbij, voor twee willekeurige possibilistische variabelen X_1 en X_2:

$$E_{\mathcal{X}}(X_1, X_2) = \widetilde{p}_{(X_1, X_2)} \tag{2.47}$$

waarbij:

$$\mu_{\widetilde{p}_{(X_1,X_2)}}(T) = \sup_{(o_1,o_2)\in O^2} \min\left(\pi_{X_1\times X_2}(o_1, o_2), \mu_{E_O(o_1,o_2)}(T)\right)$$

$$\mu_{\widetilde{p}_{(X_1,X_2)}}(F) = \sup_{(o_1,o_2)\in O^2} \min\left(\pi_{X_1\times X_2}(o_1, o_2), \mu_{E_O(o_1,o_2)}(F)\right)$$

en

$$\pi_{X_1\times X_2}(o_1, o_2) = t\left(\pi_{X_1}(o_1), \pi_{X_2}(o_2)\right) \tag{2.48}$$

onder de voorwaarde dat X_1 en X_2 t-onafhankelijk zijn.

Definitie 2.19 is intuïtief aanvaardbaar. Inderdaad, indien het volledig mogelijk is dat variabele $X_1 \times X_2$ een waarde (o_1, o_2) aanneemt en het is volledig mogelijk dat deze waarden (niet) coreferent zijn, dan is het volledig mogelijk dat de possibilistische variabelen (niet) coreferent zijn. Door Definitie 2.19 kunnen we, telkens wanneer een eigenschap van een entiteit moeilijk meetbaar is, de meetresultaten vergelijken met een evaluator $E_{\mathcal{X}}$. Merken we op dat deze oplossing in het geval van niet-meetbaarheid leidt tot de resultaten in Tabel 2.1, onder de redelijke veronderstelling dat er in elk universum minstens twee objecten bestaan die met zekerheid niet coreferent zijn. In het geval van gewone (d.i. meetbare) eigenschappen, reduceert evaluator $E_{\mathcal{X}}$ tot de onderliggende evaluator E_O. Ter afsluiting van de bespreking van moeilijk-meetbare eigenschappen geven we nog de volgende stelling.

Stelling 2.2
Voor een willekeurig objectuniversum O is het resultaat van $E_{\mathcal{X}}$ steeds genormaliseerd.

Bewijs. Aangezien voor elke possibilistische variabele X de possibiliteitsverdeling π_X genormaliseerd is, geldt er dat:

$$\forall(X_1, X_2) \in \mathcal{X}^2 : \exists(o_1, o_2) \in O^2 : \pi_{X_1 \times X_2}(o_1, o_2) = 1. \tag{2.49}$$

Voor een dergelijk koppel van objecten weten we dat $E_O(o_1, o_2)$ een genormaliseerde possibiliteitsverdeling over $\mathbb{B}$ is. Bijgevolg geldt er dat:

$$\max\left(\mu_{\widetilde{p}_{(X_1,X_2)}}(T), \mu_{\widetilde{p}_{(X_1,X_2)}}(F)\right) = 1. \tag{2.50}$$

□

2.6 Semantische evaluatie

We behandelen nu de klasse van semantische evaluatoren die reeds kort vermeld zijn in het voorgaande. Semantische evaluatoren gebruiken een vooraf gedefinieerde binaire relatie over het universum O in hun bepaling van onzekerheid. Dit is in contrast met syntactische evaluatoren, die hun kennis niet specificeren als een binaire relatie, maar als een collectie van kennisregels die een uitspraak doen op basis van syntactische verbanden tussen objecten. Deze kennisregels kunnen evenwel een binaire relatie over O induceren, maar deze relatie wordt niet als dusdanig gespecificeerd door syntactische evaluatoren.

We behandelen semantische evaluatoren in dit hoofdstuk omwille van de algemeenheid van de problematiek. Syntactische evaluatoren worden in dit hoofdstuk niet verder behandeld omdat ze typisch een welbepaalde oorzaak van onzekerheid trachten op te vangen. Om die reden worden syntactische evaluatoren afzonderlijk onderzocht in Hoofdstukken 4 en 6.

We duiden nogmaals op het verschil tussen de problematiek van semantische evaluatie en de problematiek van moeilijk-meetbare eigenschappen en/of entiteiten. Enerzijds heeft moeilijk-meetbaarheid tot gevolg dat het antwoord op een meetvraag niet precies gekend is, waardoor een aantal alternatieven wordt bijgehouden. Echter, slechts één van deze alternatieven is correct. De problematiek die we anderzijds met semantische evaluatie wensen aan te pakken, houdt in dat meerdere correcte antwoorden mogelijk zijn op eenzelfde vraag. Antwoorden kunnen daarbij verschillen in specificiteit, subjectiviteit en dergelijke meer. Bovendien gaan we er bij semantische evaluatie van uit dat een syntactisch verband vinden tussen de verschillende waarden niet altijd eenvoudig of soms zelfs onmogelijk is. Eerder wordt bij deze klasse van evaluatoren vertrokken van een binaire relatie R over het universum in kwestie. Er wordt van deze relatie verondersteld dat ze een verband tussen objecten aangeeft, dat een indicatie is voor het coreferent zijn van de objecten. Zo kunnen we bijvoorbeeld een orderelatie op geografische gebieden gebruiken om te detecteren dat New York en Manhattan mogelijks coreferent zijn. De binaire relatie geeft hierbij aan dat Manhattan een deelgebied is van New York en dit vormt op zich een indicatie voor de coreferentie ervan. Merk op dat de gebruikte binaire relatie geen rechtstreekse mogelijkheid op coreferentie specificeert. Ze

geeft enkel een verband tussen twee objecten. De vraag die we hier wensen te beantwoorden luidt hoe de kennis die een binaire relatie geeft over objecten, kan worden vertaald naar een goede onzekerheidsbepaling voor coreferentie van objecten.

We noemen de te onderzoeken klasse van evaluatoren 'semantisch' omdat de binaire relatie R in veel gevallen een semantisch verband tussen objecten uitdrukt. Dit verband is syntactisch gezien niet eenvoudig of onmogelijk op te sporen. Zo is er geen syntactisch verband tussen Manhattan en New York, maar wel tussen New York en New York City. De klasse van semantische evaluatoren wordt echter niet beperkt tot het gebruik van semantische verbanden. Bepaalde syntactische verbanden kunnen bijvoorbeeld ook als een relatie worden uitgedrukt, waarvan het meest triviale voorbeeld de gelijkheidsrelatie is. Dit toont meteen aan dat sommige evaluatoren zowel tot de klasse van semantische als tot de klasse van syntactische evaluatoren behoren. Men zou als alternatieve naam relationele evaluatoren kunnen gebruiken, hetgeen de lading misschien beter dekt. We kiezen hiervoor niet om de link met semantische similariteit te bewaren en aan te tonen dat het raamwerk van algemene binaire relaties dat we hier aanreiken ruimer is dan hetgeen reeds voorhanden is in de literatuur. Laat ons beginnen met een bondig overzicht te geven van de oplossingen die bestaan om semantische verbanden te vertalen naar similariteit. Twee pioniers op dit gebied zijn Collins en Quillian ([48, 49, 50]). Voortbouwend op studies uit de psychologie, waaruit blijkt dat in het menselijk brein een onderscheid bestaat tussen het episodisch en het semantisch geheugen, stellen Collins en Quillian een netwerkmodel voorop om het semantisch geheugen te modelleren. Dit netwerk wordt voorgesteld als een gewogen graaf, waarbij de gewichten gekoppeld aan takken similariteiten uitdrukken tussen de knopen aan de beide uiteinden van de tak. Elke knoop komt in een dergelijke graaf overeen met een bepaald semantisch concept. In navolging van Collins en Quillian zijn verschillende andere methoden bedacht om de werking van het semantisch geheugen te modelleren, waaronder associatieve modellen [51, 52] en meer recent ook statistische modellen [53, 54, 55]. Steunend op het netwerkmodel zijn verschillende voorstellen gedaan om, gegeven een semantisch netwerk, similariteiten te berekenen tussen concepten. Een strekking is deze van Rada, die stelt dat semantische verbanden kunnen worden voorgesteld door enkel de "is-een" relatie te beschouwen [56, 57]. In de praktijk spelen echter andere relaties zoals "is-een-deel-van" eveneens een belangrijke rol. Meer geavanceerde methoden veronderstellen bijkomende statistische informatie over de concepten [58, 59].

In onze aanpak van semantische evaluatie staan twee zaken centraal. Eerst en vooral willen we een vertaling geven van de kennis die een binaire relatie R over O biedt, naar kennis over de binaire coreferentierelatie ($\leftrightarrow$) over O. Dit leidt tot de definitie van semantische evaluatoren. Hierbij houden we rekening met een aantal beperkingen opdat de vertaling van kennis op een intuïtieve manier zou gebeuren. Een tweede aspect is de binaire relatie R zelf. In vele gevallen wordt deze als gekend ondersteld. Echter, dit leidt naar onze mening tot de belangrijke beperking dat de relatie voorbehouden is tot het uitdrukken

van vooraf gekende (semantische) verbanden. We zullen een algemene aanpak nastreven waarbij verbanden niet op voorhand gekend of intuïtief moeten zijn. Dit heeft als gevolg dat de binaire relatie R eerst moet worden opgebouwd. Voor de constructie van deze binaire relatie kan echter geen trainingscollectie worden gebruikt. Het is namelijk zo dat kennis over de relatie R die geleerd wordt uit een trainingscollectie in het algemeen niet geëxtrapoleerd kan worden naar het volledige universum. We zullen de relatie R dus dynamisch moeten construeren, tijdens het proces van coreferentiedetectie. Het zal blijken dat een dergelijke dynamische constructie een aantal interessante voordelen met zich meebrengt.

Een eerste voordeel is dat het niet langer nodig is om op voorhand een ontologie of taxonomie te voorzien voor elk universum waarvoor semantische verbanden in rekening moeten worden gebracht. Dit vereist namelijk manuele tussenkomst in het bepalen van coreferentie. Uiteraard bestaan in de literatuur methoden voor het automatisch extraheren van ontologieën uit teksten, maar dat gegeven verandert niets aan de context-afhankelijkheid ervan. Bovendien stellen we vast dat bij nagenoeg alle taalverwerkingstechnieken een tekst wordt voorgesteld als een vector van woorden, hetgeen een beperkt taalmodel is. Voor een bespreking van enerzijds het vermijden van ontologieën en anderzijds het gebruik van een beter taalmodel, verwijzen we naar Hoofdstuk 8, waar verder wordt ingegaan op coreferentie van tekstuele documenten.

Een tweede voordeel van onze aanpak is dat het dynamisch opbouwen van een binaire relatie voor objecten de mogelijkheid biedt om welgekende relaties te herkennen. In sommige situaties is het wel degelijk mogelijk om een dynamisch opgebouwde relatie te extrapoleren naar het volledige universum. Hierdoor ontstaat de automatische generatie van wat in de literatuur bekend staat als filtereigenschappen, d.i. eigenschappen van entiteiten waarvoor (1) de overeenkomst meestal geen uitsluitsel biedt over de gelijkheid van entiteiten, (2) de zoekruimte aanzienlijk wordt beperkt en (3) de vergelijking kost-efficiënt kan gebeuren in termen van complexiteit.

Een derde en laatste voordeel van de dynamische constructie van een binaire relatie R, is dat deze relatie onmiddellijk kan worden gebruikt in het samenvoegen van coreferente objecten. Voor een verdere bespreking van dit laatste punt wordt verwezen naar Hoofdstuk 9.

Veronderstel een universum O en laat R een binaire relatie zijn over O. Een interessant gegeven is nu dat Definitie 2.10 stelt dat $\leftrightarrow$ eveneens een, weliswaar onbekende, binaire relatie over O is. Onmiddellijk stelt zich dan de vraag hoe de gekende relatie R ons kennis biedt over de relatie $\leftrightarrow$. Hiervoor moeten we de veronderstelling maken dat R een relatie is die evidentie biedt voor coreferentie. Dit komt omdat we, gegeven een binaire relatie R, alleen onderscheid kunnen maken tussen evidentie voor coreferentie en niet-coreferentie op basis van het al dan niet gerelateerd zijn van objecten via R. We veronderstellen verder dat R volledige kennis biedt, hetgeen expliciet een drastische dualiteitsvoorwaarde oplegt in de zin dat niet-gerelateerd zijn van twee objecten evidentie geeft voor het niet-coreferent zijn van deze objecten. Een semantische evaluator

zal bijgevolg beslissen dat twee objecten niet coreferent kunnen zijn wanneer ze niet verbonden zijn door R. Het wel verbonden zijn door R impliceert volledige mogelijkheid van coreferentie, maar nog geen zekerheid. Dit leidt tot de volgende definitie.

Definitie 2.20 (Semantische evaluator)
Beschouw een eindig universum O en een binaire relatie R over O. Een semantische evaluator $E_{O,R}$ is gedefinieerd als:

$$E_{O,R} : O^2 \to \mathcal{F}(\mathbb{B}) \tag{2.51}$$

zodat:

$$E_{O,R}(o_1, o_2) = \begin{cases} (1,0) & \textbf{\textit{als}} \quad o_1 = o_2 \\ \left(1, \mu_{\widetilde{p}_{(o_1,o_2)}}(F)\right) & \textbf{\textit{als}} \quad o_1 \; R \; o_2 \vee o_2 \; R \; o_1 \\ (0,1) & \textbf{\textit{anders}}. \end{cases} \tag{2.52}$$

Een semantische evaluator is volgens Definitie 2.20 dus altijd reflexief en symmetrisch, ongeacht R. De toekenning van de possibiliteit voor niet-coreferentie (d.i. $\mu_{\widetilde{p}_{(o_1,o_2)}}(F)$) in het geval waarbij twee objecten deel uitmaken van de relatie R, is onderworpen aan voorwaarden. Vooreerst kunnen verschillende binaire relaties een verschillende informatiewaarde bevatten. Gelet op het feit dat het aantal koppels van coreferente objecten in de praktijk zeer klein is in vergelijking met het aantal niet-coreferente objecten, stellen we hier de kardinaliteit van een relatie voorop als een maat voor haar informatieve waarde. Als een groot aantal objecten uit het eindige universum verbonden is door de binaire relatie R, dan is die relatie R weinig specifiek. De relatie maakt weinig onderscheid tussen koppels van objecten en ze biedt bijgevolg weinig informatie. Deze vaststelling vertaalt zich naar de volgende vereiste:

$$R_1 \subseteq R_2 \Rightarrow \forall (o_1, o_2) \in R_1 : E_{O,R_1}(o_1, o_2) \leq E_{O,R_2}(o_1, o_2). \tag{2.53}$$

In het extreme geval waarbij een binaire relatie R over een universum O gelijk is aan $O \times O$, biedt ze totaal geen informatie. In dit geval moet gelden dat:

$$\forall (o_1, o_2) \in O^2 : (o_1 \neq o_2) \Rightarrow E_{O,R}(o_1, o_2) = (1,1). \tag{2.54}$$

Stelling 2.3
Voorwaarden (2.53) *en* (2.54) *zijn voldaan als geldt:*

$$\forall (o_1, o_2) \in R : E_{O,R}(o_1, o_2) = \left(1, g\left(\frac{|\operatorname{sel}_{o_1}(R)| - 1}{2|O| - 2}, \frac{|\operatorname{sel}_{o_2}(R)| - 1}{2|O| - 2}\right)\right) \tag{2.55}$$

onder de voorwaarde dat $o_1 \neq o_2$. Verder is $g : [0,1]^2 \to [0,1]$ een stijgende, symmetrische functie zodat $g(1,1) = 1$ en waarbij:

$$\operatorname{sel}_o(R) = \{(o_1, o_2) | o_1 \; R \; o_2 \wedge (o = o_1 \vee o = o_2)\}. \tag{2.56}$$

Bewijs. Als $R_1 \subseteq R_2$ dan geldt er:

$$\forall o \in O : \mathrm{sel}_o(R_1) \subseteq \mathrm{sel}_o(R_2) \tag{2.57}$$

en dus ook:

$$\forall o \in O : |\mathrm{sel}_o(R_1)| \leq |\mathrm{sel}_o(R_2)|. \tag{2.58}$$

Aangezien g stijgend is, geldt (2.53). Wanneer $R = O^2$, dan geldt:

$$\forall o \in O : |\mathrm{sel}_o(R)| = 2|O| - 1 \tag{2.59}$$

zodat (2.54) eveneens geldt. □

Een semantische evaluator die voldoet aan de voorwaarde uit Stelling 2.3 wordt *kardinaliteitsgebaseerd* genoemd. In het vervolg veronderstellen we dat een semantische evaluator *kardinaliteitsgebaseerd* is. Voor *kardinaliteitsgebaseerde* evaluatoren geldt de volgende stelling.

Stelling 2.4
Als $E_{O,R}$ een kardinaliteitsgebaseerde *semantische evaluator is, dan geldt voor elk triplet $(o_1, o_2, o_3) \in O^3$ van verschillende objecten waarvoor:*

$$(o_1 R o_3 \vee o_3 R o_1) \wedge (o_2 R o_3 \vee o_3 R o_2) \tag{2.60}$$

dat:

$$|\mathrm{sel}_{o_1}(R)| \geq |\mathrm{sel}_{o_2}(R)| \Rightarrow E_{O,R}(o_1, o_3) \leq E_{O,R}(o_2, o_3) \tag{2.61}$$

Bewijs. Aangezien zowel o_1 als o_2 verbonden zijn aan o_3 door R geldt:

$$E_{O,R}(o_1, o_3) = \left(1, g\left(\frac{|\mathrm{sel}_{o_1}(R)| - 1}{2|O| - 2}, \frac{|\mathrm{sel}_{o_3}(R)| - 1}{2|O| - 2}\right)\right) \tag{2.62}$$

$$E_{O,R}(o_2, o_3) = \left(1, g\left(\frac{|\mathrm{sel}_{o_2}(R)| - 1}{2|O| - 2}, \frac{|\mathrm{sel}_{o_3}(R)| - 1}{2|O| - 2}\right)\right). \tag{2.63}$$

Gelet op het feit dat g stijgend is, volgt het gestelde. □

Transitieve relaties spelen een bijzondere rol met betrekking tot semantische evaluatoren. Dit wordt bekrachtigd door de volgende stelling.

Stelling 2.5
Laat $E_{O,R}$ een kardinaliteitsgebaseerde *semantische evaluator zijn. Als R een equivalentierelatie is, dan is $E_{O,R}$ een transitieve evaluator.*

Bewijs. Als R een equivalentierelatie is, induceert R een partitie van O in m klassen K_j. Voor twee objecten o_1 en o_2 binnen dezelfde klasse geldt er dat:

$$\mathrm{sel}_{o_1}(R) = \mathrm{sel}_{o_2}(R). \tag{2.64}$$

Dit betekent dat er voor twee verschillende objecten o_1 en o_2 binnen de klasse K_j geldt dat:

$$E_{O,R}(o_1, o_2) = \widetilde{p}_j \tag{2.65}$$

met $\widetilde{p}_j$ een constante possibilistische waarheidswaarde voor klasse K_j. Gelet op het feit dat er voor twee objecten o_1 en o_2 uit een verschillende klasse geldt:

$$E_{O,R}(o_1, o_2) = (0, 1) \tag{2.66}$$

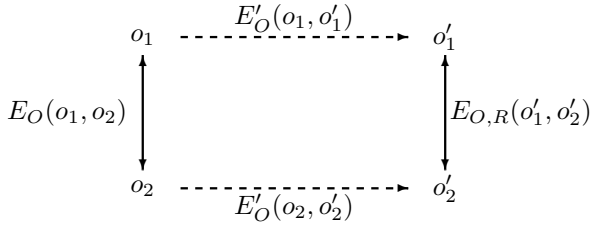
is de transitiviteit van $E_{O,R}$ bewezen. □

Zoals eerder gezegd, wordt in vele gevallen de relatie R als gekend verondersteld. Hoewel reeds vermeld is dat we in onze aanpak algemener te werk willen gaan, merken we op dat een dergelijke aanpak met een bijzonder probleem kan kampen. Meer bepaald is het zo dat vooraf bepaalde relaties uitgaan van standaardnotaties, daar waar het gebruik van standaarden bij de creatie van objecten niet wordt gegarandeerd. In een ontologie voor geografische gebieden zullen we dus "Manhattan" terugvinden, maar niet "Manhattn" omwille van de schrijffout. In zulke gevallen kan de extractie van kennis uit R worden bemoeilijkt. Dit probleem kan worden opgelost door een tweede (syntactische) evaluator over O te veronderstellen, die coreferenties aangeeft tussen geobserveerde objecten en opgeslagen standaardobjecten. Er wordt gezegd dat de semantische evaluator opgenomen is in een evaluatieketting.

Definitie 2.21 (Evaluatieketting)
Gegeven een universum O waarover een semantische evaluator $E_{O,R}$ en een evaluator E'_O gedefinieerd zijn. Een evaluatieketting wordt gevormd door een evaluator E_O zodat:

$$E_O(o_1, o_2) = \max_{(o'_1, o'_2) \in O^2} \left(E'_O(o_1, o'_1) \mathbin{\widetilde{\wedge}} E_{O,R}(o'_1, o'_2) \mathbin{\widetilde{\wedge}} E'_O(o'_2, o_2) \right). \tag{2.67}$$

Het principe van een evaluatieketting wordt getoond in Figuur 2.4.



Figuur 2.4: Evaluatieketting

We onderzoeken nu hoe een relatie R dynamisch kan worden opgebouwd. Laat ons een universum $O = U_1 \times ... \times U_n$ veronderstellen, waarbij er voor elk universum U_i een evaluator E_{U_i} is voorzien. Veronderstel dat er voor een universum U_m ($m \in \{1, ..., n\}$) een semantische evaluator moet worden voorzien. Het geval waarbij meerdere semantische evaluatoren gebruikt moeten worden is volledig analoog. Laat ons deze evaluator expliciet noteren als $E_{U_m,R}$. Indien R onbekend is, dan kan deze worden opgebouwd tijdens het

zoeken naar coreferente objecten in O. Voor de andere universa zijn syntactische evaluatoren voorhanden, waarbij elke evaluator E_{U_i} objecten vergelijkt in U_i. In Hoofdstukken 6 en 7 zal worden uitgelegd hoe op basis van een evaluator een kandidaatverzameling $\widehat{C}_i$ kan worden afgeleid. Door gebruik van de kandidaatverzamelingen van syntactische evaluatoren kan een multiverzameling $M \in \mathcal{M}(U_m \times U_m)$ worden geconstrueerd zodat $\omega_M(u_1, u_2)$ gelijk is aan:

$$\sum_{j \neq m} \left| \left\{ (o_1, o_2) | (o_1, o_2) \in \widehat{C}_j \wedge \text{proj}_m(o_1) = u_1 \wedge \text{proj}_m(o_2) = u_2 \right\} \right| \tag{2.68}$$

Door een k-snede te nemen van deze multiverzameling wordt vereist dat een koppel van deelobjecten (u_1, u_2) minstens k keer voorkomt in het geheel van kandidaatverzamelingen. Elke snede levert dus een relatie R op, die door de semantische evaluator $E_{U_m,R}$ kan worden gebruikt. Hoe groter de drempelwaarde k die wordt gebruikt voor de snede, hoe kleiner de relatie R wordt. We zullen dit toepassen in Hoofdstuk 7.

Een dergelijke aanpak heeft verschillende voordelen. Eerst en vooral wordt R dynamisch geconstrueerd, waardoor R niet noodzakelijk een gekend verband moet uitdrukken, maar eerder een verband dat representatief is voor het coreferent zijn van objecten. Ten tweede kan een dergelijke evaluator gebruikt worden als een filter met lage complexiteit. Meer bepaald worden semantische evaluatoren typisch gebruikt voor een universum met een lage kardinaliteit, d.i. weinig verschillende objecten. Dit maakt het makkelijk om een representatieve steekproef uit een datacollectie te nemen. Als op deze steekproef een relatie R geconstrueerd wordt die de extrapolatie-eigenschap bezit, dan kan R als een filter worden gebruikt om objecten te selecteren die mogelijks coreferent zijn. Meer bepaald, twee objectkoppels worden niet geselecteerd voor vergelijking (weggefilterd) als hun projecties voor universum U_m niet gerelateerd zijn via R. Een dergelijk filter is nuttig voor grote databanken met miljoenen records, waarbij de kwadratische complexiteit van coreferentiedetectie een groot probleem vormt. Het zal worden aangetoond in Hoofdstuk 9 dat een binaire relatie R zoals hier bepaald bijzonder nuttig is tijdens het samenvoegen van coreferente objecten.

2.7 Gedeeltelijke coreferentie

Hoewel de nadruk van deze thesis ligt op het coreferentieprobleem, bestaan een reeks verwante problemen die we samenvatten onder de noemer 'gedeeltelijke coreferentie'. In een meer algemene context kunnen we namelijk een veralgemeende referentiefunctie ρ^* beschouwen.

Definitie 2.22 (Veralgemeende objectreferentie)
Voor een referentiefunctie ρ is een veralgemeende referentiefunctie ρ^* een functie:

$$\rho^* : O \to \mathcal{P}(\mathcal{E}) \tag{2.69}$$

Objecten die op basis van een veralgemeende referentiefunctie ρ^* een collectie van entiteiten beschrijven, kunnen worden opgesplitst in twee klassen. De eerste klasse beslaat objecten o die een collectie van entiteiten beschrijven, zodat elke entiteit e binnen deze collectie door een afzonderlijk object wordt beschreven. Een object o is dan een multiverzameling van elementaire objecten, waarbij elk van deze elementaire objecten precies één entiteit beschrijft. In dit geval kan de beschrijving van een collectie van entiteiten door een object o volledig equivalent worden gekarakteriseerd door de verschillende elementaire objecten en de referentiefunctie ρ. Beschouw het volgende voorbeeld:

$$o_1 = \{\text{“Bijloke concertzaal”}, \text{“Bijloke museum”}\}$$

Het object o_1 bevat twee elementaire objecten die elk een afzonderlijke entiteit beschrijven. De problematiek omtrent deze klasse van objecten wordt besproken in Hoofdstuk 4. De tweede klasse beslaat objecten o die een collectie van entiteiten beschrijven als een atomair geheel, zonder enige garantie dat de afzonderlijke entiteiten identificeerbaar zijn. Beschouw als voorbeeld hiervan het object:

$$o_2 = \text{“Bijloke”}$$

Het object o_2 refereert nu naar een collectie van entiteiten die de entiteiten beschreven door o_1 bevat. Het is zo dat o_2 niet coreferent is met één van beide elementaire objecten uit o_1. Ook kunnen de entiteiten beschreven door de deelobjecten van o_1 niet worden herkend in o_2. Hoewel deze objecten dus niet onderling coreferent zijn, kan het nuttig zijn deze aan elkaar te linken, bijvoorbeeld om redenen van samenvoeging (Hoofdstuk 9). Indien een dergelijke noodzaak zich opdringt, kunnen relaties verwant aan $\leftrightarrow$ worden gedefinieerd.

Definitie 2.23 ($\subseteq$-coreferentie van objecten)
Gegeven een universum O en een veralgemeende referentiefunctie ρ^*. Twee objecten $(o_1, o_2) \in O^2$ zijn $\subseteq$-coreferent, genoteerd als $o_1 \leftrightarrow_{\subseteq} o_2$ als:

$$\rho^*(o_1) \subseteq \rho^*(o_2). \tag{2.70}$$

Definitie 2.24 ($\cap$-coreferentie van objecten)
Gegeven een universum O en een veralgemeende referentiefunctie ρ^*. Twee objecten $(o_1, o_2) \in O^2$ zijn $\cap$-coreferent, genoteerd als $o_1 \leftrightarrow_{\cap} o_2$ als:

$$\rho^*(o_1) \cap \rho^*(o_2) \neq \emptyset. \tag{2.71}$$

We kunnen inzien dat $\leftrightarrow_{\subseteq}$ een partiële orderelatie over O is (reflexief en transitief), terwijl $\leftrightarrow_{\cap}$ een afhankelijkheidsrelatie over O is (reflexief en symmetrisch). Beide relaties voldoen slechts gedeeltelijk aan de eigenschappen van $\leftrightarrow$, waardoor we spreken van gedeeltelijke coreferentie. De relatie $\leftrightarrow_{\subseteq}$ is nuttig wanneer entiteiten niet strikt onafhankelijk zijn van elkaar, maar eerder een hiërarchische structuur vormen. Deze hiërarchische structuur modelleren we dan door te stellen dat een entiteit steeds samengesteld is uit entiteiten die lager in de structuur voorkomen, waardoor het gebruik van $\leftrightarrow_{\subseteq}$ een goed model

vormt voor dit probleem. De problematiek komt hier voort uit een verschillend niveau van specificiteit dat wordt gebruikt door verschillende meetprocessen. De relatie $\leftrightarrow_{\cap}$ kan worden toegepast bij problemen waar het niet altijd duidelijk is naar welke entiteit wordt gerefereerd. Dit is bijvoorbeeld typisch het geval bij het zoeken van coreferente teksten, een probleem dat wordt besproken in Hoofdstuk 8.

Ondanks de verschillende definities kan voor deze problemen steeds een gelijkaardige aanpak worden gebruikt. De bespreking van dergelijke methoden wordt gevoerd in Hoofdstuk 5. In wat volgt zal echter systematisch uitgegaan worden van de relatie $\leftrightarrow$, tenzij expliciet anders vermeld.

2.8 Conclusie

In dit hoofdstuk is eerst uiteengezet hoe de problematiek van coreferentie tot stand komt, dit door objectcreatie te zien als een meetproces. Vervolgens is een evaluator gedefinieerd als een operator die twee objecten vergelijkt en de mogelijkheid van (niet)-coreferentie uitdrukt. Vermits de coreferentierelatie een equivalentierelatie is, is bestudeerd hoe de eigenschappen van een dergelijke relatie impact hebben op een evaluator. Verder is beschreven hoe een evaluator omgaat met moeilijk-meetbare eigenschappen van entiteiten. Ten slotte is het geval van semantische evaluatoren bestudeerd, waarbij onderzocht is hoe een binaire relatie kan worden gebruikt om kennis over coreferentie af te leiden. In het bijzonder is een techniek voorgesteld waarbij R niet op voorhand gekend hoeft te zijn. Een dergelijke techniek heeft een aantal interessante voordelen, zoals het efficiënt construeren van filters.

Hoofdstuk 3

Combinatie van onzekerheid over Boolese proposities

3.1 Inleiding

Na het invoeren van de definitie van een evaluator en het bespreken van semantische evaluatoren in het vorige hoofdstuk, is het wenselijk syntactische evaluatoren te definiëren. Het zal later blijken dat de combinatie van onzekerheid over Boolese proposities een essentiële bouwsteen is voor de constructie van dergelijke evaluatoren. Daarom wordt de combinatie van onzekerheid in dit hoofdstuk onderzocht. Hiervoor wordt in Sectie 3.2 een studie gemaakt van bestaande technieken voor combinatie van onzekerheid, waarbij de inleidende begrippen uit Hoofdstuk 1 veelvuldig worden gebruikt. Na deze studie zal worden geargumenteerd waarom een nieuwe methode nodig is. Een dergelijke methode wordt daarna opgebouwd in Sectie 3.3 en de eigenschappen worden uitgebreid besproken. We zullen in onze aanpak bestuderen hoe kennis in de vorm van possibilistische waarheidswaarden tot stand komt. Daarna construeren we een raamwerk van conditionele necessiteit, hetgeen leidt tot een nieuwe combinatiefunctie. Deze functie noemen we de Sugeno-integraal voor possibilistische waarheidswaarden. We bestuderen de eigenschappen van deze combinatiefunctie in het kader van het praktische gebruik ervan. In Sectie 3.4 worden de belangrijkste resultaten uit dit hoofdstuk samengevat.

3.2 Overzicht van de literatuur

Zoals tot hiertoe steeds is verondersteld, wordt onzekerheid over een Boolese propositie gecodeerd in de vorm van een possibilistische waarheidswaarde (Definitie 1.7). In navolging van Dubois en Prade, kan dit worden gezien als de kennis die een actor heeft over het waar of vals zijn van een propositie. Echter, tot hiertoe is nog niets gezegd over situaties waarin de kennis van verschillende

actoren moet worden gecombineerd. In dergelijke gevallen is er nood aan een mechanisme waarmee kennis in de vorm van possibiliteit kan worden gecombineerd. Gelet op de Boolese aard van dit probleem, denken we in de eerste plaats aan Boolese functies, in het bijzonder aan conjunctie ($\wedge$) en disjunctie ($\vee$). De veralgemening van deze Boolese functies is dan ook een goed bestudeerd onderwerp in de literatuur, waarbij verschillende mogelijkheden zijn onderzocht.

Beschouwen we het Boolese domein $\mathbb{B} = \{T, F\}$, dat een complete tralie is met grootste element T (waar) en kleinste element F (vals). De orderelatie op dit domein is bijgevolg gegeven door $F \leq T$. De Boolese conjunctie $\wedge$ en disjunctie $\vee$ zijn twee functies gedefinieerd als volgt.

Definitie 3.1 (Conjunctie $\wedge$)
Conjunctie van twee Boolese variabelen p en q is gedefinieerd als:

$$\wedge : \mathbb{B}^2 \to \mathbb{B} : (p, q) \mapsto p \wedge q \tag{3.1}$$

waarbij:

$$p \wedge q = \begin{cases} T & \textbf{als} \quad p = q = T \\ F & \textbf{anders}. \end{cases} \tag{3.2}$$

Definitie 3.2 (Disjunctie $\vee$)
Disjunctie van twee Boolese variabelen p en q is gedefinieerd als:

$$\wedge : \mathbb{B}^2 \to \mathbb{B} : (p, q) \mapsto p \vee q \tag{3.3}$$

waarbij:

$$p \wedge q = \begin{cases} F & \textbf{als} \quad p = q = F \\ T & \textbf{anders}. \end{cases} \tag{3.4}$$

Gelet op de orderelatie $\leq$ over $\mathbb{B}$ stellen we vast dat $\wedge$ en $\vee$ respectievelijk de minimumoperator min en de maximumoperator max voor het domein $\mathbb{B}$ zijn. Om die reden zullen we het symbool $\wedge$ ($\vee$) gebruiken om de operator min (max) aan te duiden op zowel het domein $\mathbb{B}$ als $\mathbb{R}$. Aangezien possibilistische waarheidswaarden genormaliseerde vaagverzamelingen over het domein $\mathbb{B}$ zijn, kunnen $\wedge$ en $\vee$ worden uitgebreid, zodat vage functies in het domein van possibilistische waarheidswaarden worden verkregen. Gebruiken we hiervoor het uitbreidingsprincipe van Zadeh (Definitie 1.5) dan krijgen we volgende definities.

Definitie 3.3 (Zadeh-uitbreiding van $\wedge$)
Gegeven twee possibilistische waarheidswaarden $\widetilde{p}$ en $\widetilde{q}$, dan is de Zadeh-uitbreiding van $\wedge$ gedefinieerd als:

$$\widetilde{\wedge} : \mathcal{F}(\mathbb{B})^2 \to \mathcal{F}(\mathbb{B}) : (\widetilde{p}, \widetilde{q}) \mapsto \widetilde{p} \,\widetilde{\wedge}\, \widetilde{q} \tag{3.5}$$

waarbij:

$$\mu_{\widetilde{p}\ \widetilde{\wedge}\ \widetilde{q}}(T) = \min\left(\mu_{\widetilde{p}}(T), \mu_{\widetilde{q}}(T)\right) \tag{3.6}$$

$$\mu_{\widetilde{p}\ \widetilde{\wedge}\ \widetilde{q}}(F) = \max\begin{pmatrix} \min\left(\mu_{\widetilde{p}}(T), \mu_{\widetilde{q}}(F)\right), \\ \min\left(\mu_{\widetilde{p}}(F), \mu_{\widetilde{q}}(T)\right), \\ \min\left(\mu_{\widetilde{p}}(F), \mu_{\widetilde{q}}(F)\right) \end{pmatrix}. \tag{3.7}$$

Definitie 3.4 (Zadeh-uitbreiding van $\vee$)
Gegeven twee possibilistische waarheidswaarden $\widetilde{p}$ en $\widetilde{q}$, dan is de Zadeh-uitbreiding van $\vee$ gedefinieerd als:

$$\widetilde{\vee}\ : \mathcal{F}(\mathbb{B})^2 \to \mathcal{F}(\mathbb{B}) : (\widetilde{p}, \widetilde{q}) \mapsto \widetilde{p}\ \widetilde{\vee}\ \widetilde{q} \tag{3.8}$$

waarbij:

$$\mu_{\widetilde{p}\ \widetilde{\vee}\ \widetilde{q}}(T) = \max\begin{pmatrix} \min\left(\mu_{\widetilde{p}}(T), \mu_{\widetilde{q}}(T)\right), \\ \min\left(\mu_{\widetilde{p}}(F), \mu_{\widetilde{q}}(T)\right), \\ \min\left(\mu_{\widetilde{p}}(T), \mu_{\widetilde{q}}(F)\right) \end{pmatrix} \tag{3.9}$$

$$\mu_{\widetilde{p}\ \widetilde{\vee}\ \widetilde{q}}(F) = \min\left(\mu_{\widetilde{p}}(F), \mu_{\widetilde{q}}(F)\right). \tag{3.10}$$

Definities 3.3 en 3.4 zijn voor het eerst ingevoerd door Van Schooten [29]. In [22] beschrijft De Cooman een formele afleiding van het uitbreidingsprincipe van Zadeh. Hierbij wordt vertrokken van een formele invoering van transformatie van possibiliteitsmaten door afbeeldingen. Vanuit dit formeel standpunt definieert De Cooman het possibilistische uitbreidingsprincipe, waaruit het uitbreidingsprincipe van Zadeh voortkomt als zijnde een speciaal geval. Het uitbreidingsprincipe van De Cooman is bijgevolg algemener dan dit van Zadeh en het steunt op een volledig formele redenering. Een volledige bespreking van het werk van De Cooman ligt buiten het bestek van deze thesis. We zullen ons beperken tot zijn resultaten met betrekking tot Boolese logica. De Cooman toont aan dat een afbeelding $X_1 \times ... \times X_n \to Y$ kan worden uitgebreid volgens het possibilistische uitbreidingsprincipe, op voorwaarde dat de possibilistische variabelen $\widetilde{x}_i$ in X_i t-onafhankelijk zijn. Dit noemt De Cooman possibilistische t-uitbreidingen. Vervolgens past hij zijn resultaten toe op de klassieke Boolese logica, waardoor de operator min in Definities 3.3 en 3.4 vervangen wordt door een triangulaire norm t, gesteld dat $\widetilde{p}$ en $\widetilde{q}$ t-onafhankelijk zijn. Ook toont De Cooman aan dat in het geval van conjunctie (disjunctie), de resulterende possibiliteit voor F (T) kan worden vereenvoudigd. Dit geeft aanleiding tot volgende definities.

Definitie 3.5 (Possibilistische t-uitbreiding van $\wedge$)
Gegeven twee possibilistische waarheidswaarden $\widetilde{p}$ en $\widetilde{q}$, dan is de possibilistische t-uitbreiding van $\wedge$ gedefinieerd als:

$$\widetilde{\wedge}_t\ : \mathcal{F}(\mathbb{B})^2 \to \mathcal{F}(\mathbb{B}) : (\widetilde{p}, \widetilde{q}) \mapsto \widetilde{p}\ \widetilde{\wedge}_t\ \widetilde{q} \tag{3.11}$$

waarbij:

$$\mu_{\widetilde{p}\ \widetilde{\wedge}_t\ \widetilde{q}}(T) = t\left(\mu_{\widetilde{p}}(T), \mu_{\widetilde{q}}(T)\right) \tag{3.12}$$
$$\mu_{\widetilde{p}\ \widetilde{\wedge}_t\ \widetilde{q}}(F) = \max\left(\mu_{\widetilde{p}}(F), \mu_{\widetilde{q}}(F)\right) \tag{3.13}$$

op voorwaarde dat $\widetilde{p}$ en $\widetilde{q}$ t-onafhankelijk zijn met t een triangulaire norm.

Definitie 3.6 (Possibilistische t-uitbreiding van $\vee$)
Gegeven twee possibilistische waarheidswaarden $\widetilde{p}$ en $\widetilde{q}$, dan is de possibilistische t-uitbreiding van $\vee$ gedefinieerd als:

$$\widetilde{\vee}_t : \mathcal{F}(\mathbb{B})^2 \to \mathcal{F}(\mathbb{B}) : (\widetilde{p}, \widetilde{q}) \mapsto \widetilde{p}\ \widetilde{\vee}_t\ \widetilde{q} \tag{3.14}$$

waarbij:

$$\mu_{\widetilde{p}\ \widetilde{\vee}_t\ \widetilde{q}}(T) = \max\left(\mu_{\widetilde{p}}(T), \mu_{\widetilde{q}}(T)\right) \tag{3.15}$$
$$\mu_{\widetilde{p}\ \widetilde{\vee}_t\ \widetilde{q}}(F) = t\left(\mu_{\widetilde{p}}(F), \mu_{\widetilde{q}}(F)\right) \tag{3.16}$$

op voorwaarde dat $\widetilde{p}$ en $\widetilde{q}$ t-onafhankelijk zijn met t een triangulaire norm.

De nood aan het modelleren van afhankelijkheden kan worden aangetoond vanuit een formeel standpunt. Beschouw twee proposities p en q, waarbij $q = \neg p$, dan is $p \wedge q$ altijd gelijk aan F. Het wordt gezegd dat $p \wedge q$ een contradictie is. Echter, wanneer $\widetilde{p} = (1, 1)$ en dus ook $\widetilde{q} = (1, 1)$, dan bestaat er geen enkele triangulaire norm zodat $\widetilde{p}\ \widetilde{\wedge}_t\ \widetilde{q}$ gelijk is aan $(0, 1)$. Volgens De Cooman is het dus niet mogelijk om de conjunctie van deze twee proposities te berekenen. Deze bevinding stemt overeen met de fundamentele kennisbeschrijvende natuur van possibilistische waarheidswaarden. De berekening van $\widetilde{p}\ \widetilde{\wedge}_t\ \widetilde{q}$ steunt louter en alleen op de kennis over proposities p en q en niet op de kennis dat p en q elkaars tegengestelden zijn. Er bestaat geen triangulaire norm die deze kennis kan modelleren. We zullen verderop in dit hoofdstuk de interpretatie van t-onafhankelijkheid verder bespreken.

Tot slot vermelden we het werk van De Tré en De Baets [60], die een mechanisme invoeren om onzekerheid op een eenvoudige manier te transformeren. Op deze manier willen zij functies creëren in $\mathcal{F}(\mathbb{B})$ die geen rechtstreekse uitbreidingen zijn van $\wedge$ of $\vee$. Om hun transformaties te bewerkstelligen steunen De Tré en De Baets op residuele implicatoren en residuele co-implicatoren.

Definitie 3.7 (Residuele implicator)
Een residuele implicator (of R-implicator) is gedefinieerd als een functie:

$$f_{im,t} : [0,1]^2 \to [0,1] \tag{3.17}$$
$$(x, y) \mapsto \sup\left\{z | z \in [0,1] \wedge t\left(x, z\right) \leq y\right\} \tag{3.18}$$

met t een willekeurige triangulaire norm. Er geldt voor alle $(x, y) \in [0, 1]^2$:

$$f_{im,t}(x, y) = \begin{cases} 1 & \textbf{als} \quad x \leq y \\ t(x, y) & \textbf{anders}. \end{cases} \tag{3.19}$$

Definitie 3.8 (Residuele co-implicator)
Een residuele co-implicator (of R-co-implicator) is gedefinieerd als een functie:

$$f_{im,s}^{co} : [0,1]^2 \rightarrow [0,1] \tag{3.20}$$
$$(x,y) \mapsto \inf\{z | z \in [0,1] \wedge s(x,z) \geq y\} \tag{3.21}$$

met s een willekeurige triangulaire conorm. Er geldt voor alle $(x,y) \in [0,1]^2$:

$$f_{im,s}^{co}(x,y) = \begin{cases} 0 & \textbf{als} \quad x \geq y \\ s(x,y) & \textbf{anders}. \end{cases} \tag{3.22}$$

De functies $f_{im,t}$ en $f_{im,s}^{co}$ zijn $[0,1]$-uitbreidingen van respectievelijk de Boolese implicatie en de Boolese co-implicatie. Merk op dat wanneer $(t,s,1-.)$ een De Morgan triplet is, dan geldt er:

$$f_{im,t}(x,y) = 1 - f_{im,s}^{co}(1-x, 1-y). \tag{3.23}$$

De Tré en De Baets gebruiken deze operatoren voor transformatie van onzekerheid in de vorm van possibilistische waarheidswaarden, waarbij ze onzekerheid willen verhogen of verlagen wanneer de possibiliteiten voor T (resp. F) aan bepaalde voorwaarden voldoen. Ze argumenteren dat er situaties zijn waarbij onzekerheid genegeerd mag worden, op voorwaarde dat de bestaande onzekerheid klein genoeg is. In het duale scenario kunnen er situaties bestaan waarin onzekerheid gemaximaliseerd moet worden, op voorwaarde dat de bestaande onzekerheid groot genoeg is. Deze voorwaarden worden gecontroleerd met behulp van drempelwaarden voor de possibiliteiten. De Tré en De Baets tonen aan dat één van de argumenten van een implicator (co-implicator) fungeert als een drempelwaarde, als deze implicator (co-implicator) steunt op een idempotente triangulaire norm (conorm). Om die reden beschouwen De Tré en De Baets enkel de functies $f_{im,\min}$ en $f_{im,\max}^{co}$, die de implicator en co-implicator van Gödel worden genoemd en in het vervolg genoteerd zullen worden als f_{im} en f_{im}^{co}. Om tot transformatiefuncties voor possibilistische waarheidswaarden te komen, halen De Tré en De Baets aan dat het negeren (d.i. neutraliseren) van onzekerheid afhangt van de combinatiefunctie die zal worden toegepast op de getransformeerde possibilistische waarheidswaarden. Meer specifiek houdt neutralisatie een transformatie in naar een neutraal element. Dit neutraal element hangt af van de gebruikte combinatiefunctie. Op die manier komen De Tré en De Baets tot volgende definitie in het geval van conjunctie.

Definitie 3.9 (Conjunctieve transformatie in $\mathcal{F}(\mathbb{B})$)
De conjunctieve transformatie van possibilistische waarheidswaarden is gegeven door een functie $g_c : ([0,1] \times \mathcal{F}(\mathbb{B})) \rightarrow \mathcal{F}(\mathbb{B})$ waarvoor er geldt:

$$\mu_{g_c(x,\widetilde{p})}(T) = f_{im}(x, \mu_{\widetilde{p}}(T)) \tag{3.24}$$
$$\mu_{g_c(x,\widetilde{p})}(F) = f_{im}^{co}(1-x, \mu_{\widetilde{p}}(F)). \tag{3.25}$$

Hoewel dit in [60] niet expliciet is vermeld, kan het duale geval voor disjunctie analoog worden gedefinieerd.

Definitie 3.10 (Disjunctieve transformatie in $\mathcal{F}(\mathbb{B})$)
De disjunctieve transformatie van possibilistische waarheidswaarden is gegeven door een functie $g_d : ([0,1] \times \mathcal{F}(\mathbb{B})) \to \mathcal{F}(\mathbb{B})$ waarvoor geldt:

$$\begin{aligned} \mu_{g_d(x,\widetilde{p})}(T) &= f_{im}^{co}\left(1-x, \mu_{\widetilde{p}}(T)\right) && (3.26)\\ \mu_{g_d(x,\widetilde{p})}(F) &= f_{im}\left(x, \mu_{\widetilde{p}}(F)\right). && (3.27) \end{aligned}$$

Deze functies voldoen aan een aantal eigenschappen. Eerst en vooral zijn de neutrale elementen van de Zadeh-uitbreidingen van $\wedge$ en $\vee$ elkaars complement onder $\widetilde{\neg}$, wat in de transformatiefuncties wordt weerspiegeld als volgt:

$$\begin{aligned} g_c\left(x,\widetilde{p}\right) &= \left(f_{im}\left(x, \mu_{\widetilde{p}}(T)\right), f_{im}^{co}\left(1-x, \mu_{\widetilde{p}}(F)\right)\right)\\ &= \left(f_{im}\left(x, \mu_{\widetilde{\neg}\,\widetilde{p}}(F)\right), f_{im}^{co}\left(1-x, \mu_{\widetilde{\neg}\,\widetilde{p}}(T)\right)\right)\\ &= \widetilde{\neg}\left(f_{im}^{co}\left(1-x, \mu_{\widetilde{\neg}\,\widetilde{p}}(T)\right), f_{im}\left(x, \mu_{\widetilde{\neg}\,\widetilde{p}}(F)\right)\right)\\ &= \widetilde{\neg}\, g_d(x, \widetilde{\neg}\,\widetilde{p}). \end{aligned} \qquad (3.28)$$

Stelling 3.1
Het beeld van g_c en g_d is voor gegeven $\widetilde{p}$ beperkt tot drie verschillende waarden:

$$\begin{aligned} g_c\left(x,\widetilde{p}\right) &\in \{\widetilde{p}, (1,0), (1,1)\} && (3.29)\\ g_d\left(x,\widetilde{p}\right) &\in \{\widetilde{p}, (0,1), (1,1)\}. && (3.30) \end{aligned}$$

Bewijs. We bewijzen de stelling in het geval van g_c. Het geval van g_d is volkomen analoog. Aangezien $\widetilde{p}$ genormaliseerd is, geldt er dat:

$$\max\left(\mu_{\widetilde{p}}(T), \mu_{\widetilde{p}}(F)\right) = 1. \qquad (3.31)$$

(a) Indien $\mu_{\widetilde{p}}(T) = 1$, dan is $\mu_{g_c(x,\widetilde{p})}(T) = 1$. Bijgevolg is:

$$\left(g_c(x,\widetilde{p}) = (1,0)\right) \vee \left(g_c(x,\widetilde{p}) = (1, \mu_{\widetilde{p}}(F)) = \widetilde{p}\right). \qquad (3.32)$$

(b) Indien $\mu_{\widetilde{p}}(F) = 1$ onderscheiden we twee gevallen. Als $x = 0$ dan geldt er:

$$g_c(x,\widetilde{p}) = (1,0). \qquad (3.33)$$

Als $x > 0$, dan is $\mu_{g_c(x,\widetilde{p})}(F) = 1$. Bijgevolg is:

$$\left(g_c(x,\widetilde{p}) = (1,1)\right) \vee \left(g_c(x,\widetilde{p}) = (\mu_{\widetilde{p}}(T), 1) = \widetilde{p}\right). \qquad (3.34)$$

□

Stelling 3.1 toont aan dat beide transformatiefuncties onzekerheid neutraliseren (neutraal element), maximaliseren ($(1,1)$) of behouden ($\widetilde{p}$). Uit het bewijs van Stelling 3.1 volgt onmiddellijk dat:

$$\begin{aligned} \forall x \in [0,1] : \forall \widetilde{p} \in \mathcal{F}(\mathbb{B}) &: g_c\left(x,\widetilde{p}\right) \geq \widetilde{p} && (3.35)\\ \forall x \in [0,1] : \forall \widetilde{p} \in \mathcal{F}(\mathbb{B}) &: g_d\left(x,\widetilde{p}\right) \leq \widetilde{p}. && (3.36) \end{aligned}$$

Voorts is het zo dat:

$$\begin{aligned} \forall \widetilde{p} \in \mathcal{F}(\mathbb{B}) &: g_c\left(1, \widetilde{p}\right) = \widetilde{p} \\ \forall \widetilde{p} \in \mathcal{F}(\mathbb{B}) &: g_d\left(1, \widetilde{p}\right) = \widetilde{p} \\ \forall \widetilde{p} \in \mathcal{F}(\mathbb{B}) &: g_c\left(0, \widetilde{p}\right) = (1,0) \\ \forall \widetilde{p} \in \mathcal{F}(\mathbb{B}) &: g_d\left(0, \widetilde{p}\right) = (0,1). \end{aligned} \tag{3.37}$$

De transformatiefuncties worden door De Tré en De Baets vervolgens gebruikt in combinatie met Zadeh-uitbreidingen van $\wedge$ en $\vee$, hetgeen leidt tot volgende definities.

Definitie 3.11 (Getransformeerde Zadeh-uitbreiding van $\wedge$)
Gegeven twee possibilistische waarheidswaarden $\widetilde{p}$ en $\widetilde{q}$, dan is de getransformeerde Zadeh-uitbreiding van $\wedge$ gedefinieerd als:

$$\widetilde{\wedge^*} : \left([0,1] \times \mathcal{F}(\mathbb{B})\right)^2 \rightarrow \mathcal{F}(\mathbb{B}) : \tag{3.38}$$

$$\left((w_p, \widetilde{p}), (w_q, \widetilde{q})\right) \mapsto g_c\left(w_p, \widetilde{p}\right) \ \widetilde{\wedge} \ g_c\left(w_q, \widetilde{q}\right). \tag{3.39}$$

Definitie 3.12 (Getransformeerde Zadeh-uitbreiding van $\vee$)
Gegeven twee possibilistische waarheidswaarden $\widetilde{p}$ en $\widetilde{q}$, dan is de getransformeerde Zadeh-uitbreiding van $\vee$ gedefinieerd als:

$$\widetilde{\vee^*} : \left([0,1] \times \mathcal{F}(\mathbb{B})\right)^2 \rightarrow \mathcal{F}(\mathbb{B}) : \tag{3.40}$$

$$\left((w_p, \widetilde{p}), (w_q, \widetilde{q})\right) \mapsto g_d\left(w_p, \widetilde{p}\right) \ \widetilde{\vee} \ g_d\left(w_q, \widetilde{q}\right). \tag{3.41}$$

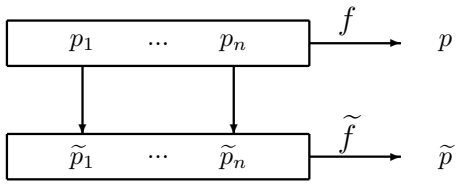
De Tré en De Baets stellen in hun werk dat $(w_p, \widetilde{p}) \ \widetilde{\wedge^*} \ (w_q, \widetilde{q})$ wel degelijk de kennis voorstelt over $p \wedge q$, maar dat de gegeven onzekerheden over deze proposities geneutraliseerd kunnen worden. Dit gegeven zal in het vervolg van dit hoofdstuk een bijzondere rol spelen.

3.3 Kenniscombinatie

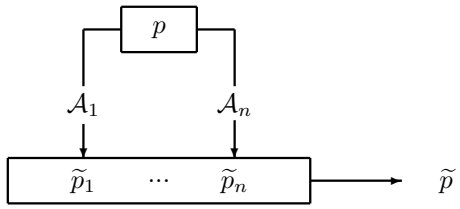
3.3.1 Een alternatieve kijk op kennisgeneratie

We zullen onze redenering omtrent kenniscombinatie beginnen met een alternatieve kijk op kennisgeneratie, d.i. de manier waarop possibilistische waarheidswaarden tot stand komen. De methoden voor kenniscombinatie die besproken zijn in Sectie 3.2 veronderstellen impliciet een model waarbij elke possibilistische waarheidswaarde onzekerheid beschrijft over de waarheidswaarde van een afzonderlijke Boolese propositie. Als dusdanig kunnen Boolese functies worden uitgebreid naar het raamwerk van possibilistische waarheidswaarden door gebruik van een uitbreidingsprincipe zoals dat van Zadeh of dat van De Cooman. Dit model wordt geïllustreerd in Figuur 3.1.

In wat volgt willen we niet langer veronderstellen dat een combinatiefunctie voor possibilistische waarheidswaarden de uitbreiding is van een logische



Figuur 3.1: Het klassieke model voor kennisgeneratie



Figuur 3.2: Het alternatieve model voor kennisgeneratie

functie uit het Boolese domein. We veronderstellen hier een context waarbij precies één propositie p is gegeven en waarbij n actoren kennis genereren over de waarheidswaarde van die propositie p. De vraag waarop we een antwoord willen vinden is dan: "Wat is de waarheidswaarde van p?". De verzameling van actoren wordt genoteerd als $A = \{\mathcal{A}_1, ..., \mathcal{A}_n\}$ waarbij $\mathcal{A}_i$ de i^{de} actor voorstelt. Elke actor formuleert een possibilistische waarheidswaarde $\widetilde{p}_i$ die de kennis van $\mathcal{A}_i$ over de waarheidswaarde van p voorstelt. De verzameling van alle possibilistische waarheidswaarden zullen we noteren als $P_\pi = \{\widetilde{p}_1, ..., \widetilde{p}_n\}$. Wanneer actor $\mathcal{A}$ beweert dat $p = T$, noteren we dit met $\mathcal{A} \rightsquigarrow (p = T)$. Wanneer actor $\mathcal{A}$ beweert dat $p = F$, noteren we dit met $\mathcal{A} \rightsquigarrow (p = F)$. Met deze notaties definiëren we voor een willekeurige $Q \subseteq A$:

$$(Q = T) \triangleq \forall \mathcal{A} \in Q : \mathcal{A} \rightsquigarrow (p = T) \tag{3.42}$$

$$(Q = F) \triangleq \forall \mathcal{A} \in Q : \mathcal{A} \rightsquigarrow (p = F)\,. \tag{3.43}$$

Onder deze definitie betekent $Q = T$ (resp. $Q = F$) dat alle actoren in Q stellen dat p waar (resp. vals) is. Voor een willekeurige $Q \subseteq A$ noteren we Q_π als de (multi)verzameling van possibilistische waarheidswaarden gegenereerd door actoren in Q. De kijk op kennisgeneratie die we hier handhaven wordt geïllustreerd in Figuur 3.2.

3.3.2 Soorten afhankelijkheden

De methode voor kenniscombinatie die we wensen in te voeren, kan worden gezien als een veralgemening van de methode van De Tré en De Baets, in die zin dat onze methode eveneens vertrekt vanuit het idee van onzekerheidsneutralisatie. Het argument om de methode van De Tré en De Baets te veralgemenen, wordt gegeven door de vaststelling dat onzekerheidsneutralisatie nuttig is in de

context van coreferentie. Stel bijvoorbeeld twee coreferente objecten waarbij de deelobjecten die de i^{de} eigenschap beschrijven, worden vergeleken door een evaluator E_i. In deze context treedt E_i bijgevolg op als een actor. Stel dat evaluator E_i postuleert dat de coreferentie van de objecten eerder onzeker is. Als er voldoende andere evaluatoren zijn die uitdrukken dat coreferentie niet onzeker is, dan is het nuttig de onzekerheid over de i^{de} eigenschap te neutraliseren. Dit verduidelijkt de noodzaak aan combinatiemethoden die een veralgemening zijn van de methode van De Tré en De Baets.

We zullen aantonen dat onze aanpak voortvloeit uit een soort van afhankelijkheid tussen actoren die orthogonaal staat op het concept 'afhankelijkheid' in de zin van De Cooman. Om verschillende soorten van afhankelijkheden te kunnen beschouwen, zullen we eerst de resultaten van De Cooman, die voortkomen uit een louter formele redenering, een interpretatie aanhechten. Laat ons daarom even terugkeren naar het klassieke model van kennisgeneratie voorgesteld in Figuur 3.1. In het geval van $\widetilde{\wedge}_t$ stelt De Cooman dat $\widetilde{p}$ en $\widetilde{q}$ t-onafhankelijk moeten zijn, opdat $\widetilde{p}\ \widetilde{\wedge}_t\ \widetilde{q}$ de kennis zou voorstellen over $p \wedge q$. Hierbij moet t-onafhankelijkheid vooraf worden verondersteld en/of geverifieerd. Aangezien de minimumoperator min puntsgewijs de grootste triangulaire norm is, geldt er dat:

$$\left(\widetilde{p}\ \widetilde{\wedge}_t\ \widetilde{q}\right) \leq \left(\widetilde{p}\ \widetilde{\wedge}\ \widetilde{q}\right). \tag{3.44}$$

Wanneer we dit vertalen naar het alternatieve model voor kennisgeneratie (Figuur 3.2) worden afhankelijkheden tussen proposities vertaald naar afhankelijkheden tussen actoren. Dit betekent dat p waar is als en alleen als alle actoren postuleren dat p waar is:

$$(p = T) \Leftrightarrow ((\mathcal{A}_1 \rightsquigarrow (p = T)) \wedge (\mathcal{A}_2 \rightsquigarrow (p = T))). \tag{3.45}$$

Gelet op (3.44) betekent een afhankelijkheid in de zin van De Cooman dat er een zekere tegenspraak is tussen actoren, waardoor p niet waar kan zijn. Afhankelijkheid in de zin van De Cooman maakt bijgevolg impliciet een veronderstelling over het soort afhankelijkheid dat er tussen actoren aanwezig is. Immers, een sterkere afhankelijkheid (d.i. een puntsgewijs kleinere triangulaire norm) tussen $\mathcal{A}_1$ en $\mathcal{A}_2$ impliceert dat de kennis over p dichter bij $(0, 1)$ ligt. Hierbij moet 'dichter' in de zin van de natuurlijke ordening over $\mathcal{F}(\mathbb{B})$ worden geïnterpreteerd (Hoofdstuk 1). Een sterkere afhankelijkheid in de zin van De Cooman reduceert dus de mogelijkheid dat p waar is. De impliciete veronderstelling die achter t-onafhankelijkheid schuilgaat, kan dus worden uitgedrukt als een vervaging van de logische implicaties:

$$(\mathcal{A}_1 \rightsquigarrow (p = T)) \quad \Rightarrow \quad (\mathcal{A}_2 \rightsquigarrow (p = F)) \tag{3.46}$$

$$(\mathcal{A}_2 \rightsquigarrow (p = T)) \quad \Rightarrow \quad (\mathcal{A}_1 \rightsquigarrow (p = F)). \tag{3.47}$$

Deze intuïtieve interpretatie kan ook worden afgeleid uit de resultaten van De Cooman zelf. Aangezien er onder de klassieke kijk op kennisgeneratie geldt dat:

$$\neg\,(p \wedge q) \equiv \neg p \vee \neg q \equiv (p \Rightarrow \neg q) \equiv (q \Rightarrow \neg p) \tag{3.48}$$

kan dit, in de veronderstelling van t-onafhankelijkheid, worden vertaald naar het universum van possibilistische waarheidswaarden:

$$\widetilde{\neg}\left(\widetilde{p}\,\widetilde{\wedge}_t\,\widetilde{q}\right) \equiv\ \widetilde{\neg}\,\widetilde{p}\,\widetilde{\vee}_t\ \widetilde{\neg}\,\widetilde{q} \equiv (\widetilde{p}\,\widetilde{\Rightarrow}_t\ \widetilde{\neg}\,\widetilde{q}) \equiv (\widetilde{q}\,\widetilde{\Rightarrow}_t\ \widetilde{\neg}\,\widetilde{p}) \tag{3.49}$$

hetgeen betekent dat t-onafhankelijkheid bepaalt hoe de kennis over $p \Rightarrow \neg q$ gemodelleerd moet worden. We zullen een dergelijke afhankelijkheid tussen actoren in het vervolg waarheidscomplementerende afhankelijkheid noemen. Dit betekent dat het waar (vals) zijn van een propositie impliceert dat een andere (afhankelijke) propositie vals (waar) is. Een volledig duale redenering kan worden gevolgd voor $\widetilde{\vee}_t$, die in het werk van De Cooman weerspiegeld wordt door:

$$(p \vee q) \equiv (\neg p \Rightarrow q) \equiv (\neg q \Rightarrow p)\,. \tag{3.50}$$

In de context waarin possibilistische waarheidswaarden hier worden gebruikt, zijn dergelijke afhankelijkheden niet nuttig. Eerder is er nood aan een aanpak waarbij we kunnen besluiten dat een propositie p waar (vals) is, zonder dat alle actoren noodzakelijk beweren dat p waar (vals) is. Anders gezegd, er is nood aan een soort van afhankelijkheid die orthogonaal staat op afhankelijkheid in de zin van De Cooman. Dergelijke afhankelijkheden nomen we waarheidsbehoudend en ze worden gegeven door de volgende implicaties:

$$(\mathcal{A}_1 \rightsquigarrow (p = T)) \quad \Rightarrow \quad (\mathcal{A}_2 \rightsquigarrow (p = T)) \tag{3.51}$$
$$(\mathcal{A}_2 \rightsquigarrow (p = T)) \quad \Rightarrow \quad (\mathcal{A}_1 \rightsquigarrow (p = T)) \tag{3.52}$$
$$(\mathcal{A}_1 \rightsquigarrow (p = F)) \quad \Rightarrow \quad (\mathcal{A}_2 \rightsquigarrow (p = F)) \tag{3.53}$$
$$(\mathcal{A}_2 \rightsquigarrow (p = F)) \quad \Rightarrow \quad (\mathcal{A}_1 \rightsquigarrow (p = F))\,. \tag{3.54}$$

Deze soort afhankelijkheid kan worden geïnterpreteerd als het vertrouwen die we in een actor stellen. Beschouwen we daartoe het volgende voorbeeld.

Voorbeeld 3.1
Beschouw de propositie p en veronderstel twee actoren, die elk kennis beschrijven over p, waarbij de kennis wordt voorgesteld door:

$$\begin{aligned} \widetilde{p}_1 &= (1, 0.3) \\ \widetilde{p}_2 &= (0.5, 1). \end{aligned}$$

Wanneer we stellen dat:

$$(\mathcal{A}_1 \rightsquigarrow (p = T)) \Rightarrow (\mathcal{A}_2 \rightsquigarrow (p = T)) \tag{3.55}$$

betekent dit dat het vertrouwen in $\mathcal{A}_1$ compleet is. Met andere woorden, als $\mathcal{A}_1$ stelt dat p waar is, dan besluiten we dat p inderdaad waar. Merk op dat als we bijkomend zouden formuleren dat:

$$(\mathcal{A}_2 \rightsquigarrow (p = F)) \Rightarrow (\mathcal{A}_1 \rightsquigarrow (p = F)) \tag{3.56}$$

dan ontstaat er een contradictie gelet op $\widetilde{p}_1$ en $\widetilde{p}_2$. We zullen in wat volgt aantonen dat onze aanpak vrij is van dergelijke contradicties.

In Voorbeeld 3.1 kan de methode van De Tré en De Baets worden gebruikt om de onzekerheid over p die wordt gepostuleerd door $\mathcal{A}_2$ te neutraliseren, mits het kiezen van de correcte drempelwaarde x. De drempelwaarde x speelt dan een analoge rol bij het modelleren van waarheidsbehoudende afhankelijkheden als die van de triangulaire norm bij het modelleren van waarheidscomplementerende afhankelijkheden. Gelet op het feit dat we de afhankelijkheden in kwestie kunnen interpreteren als vertrouwen, zullen we in de nu volgende sectie een raamwerk van conditionele necessiteit voorstellen om dit vertrouwen te modelleren.

3.3.3 Conditionele necessiteit

We zullen hier veronderstellen dat n actoren beschikbaar zijn, die elk hun kennis over een propositie p postuleren. We maken in deze sectie gebruik van de notaties die zijn ingevoerd in vorige sectie. In de context van n actoren nemen waarheidsbehoudende afhankelijkheden een algemenere vorm aan. Voor een willekeurige verzameling van actoren $Q \subset A$ kunnen de volgende afhankelijkheden bestaan:

$$(Q = T) \quad \Rightarrow \quad \left(\overline{Q} = T\right) \tag{3.57}$$
$$(Q = F) \quad \Rightarrow \quad \left(\overline{Q} = F\right). \tag{3.58}$$

Opnieuw kunnen we hier de interpretatie van vertrouwen handhaven. Onder deze interpretatie betekent (3.57) dat wanneer alle actoren in Q postuleren dat p waar is, het vertrouwen in $p = T$ volledig is. In een praktische omgeving is het niet ondenkbaar dat het vertrouwen in een groep van actoren noch 0 noch 1 is. Om die reden definiëren we twee vertrouwensmaten, die we samen de conditionele necessiteit noemen.

Definitie 3.13 (Conditionele necessiteit)
Beschouw een Boolese propositie p en een verzameling $A = \{\mathcal{A}_1, ..., \mathcal{A}_n\}$ van actoren die kennis postuleren over de waarheidswaarde van p. Het vertrouwen in deze actoren wordt bepaald door twee vertrouwensmaten γ^T en γ^F, gedefinieerd als volgt:

$$\gamma^T : \mathcal{P}(A) \to [0, 1] \tag{3.59}$$
$$\gamma^F : \mathcal{P}(A) \to [0, 1] \tag{3.60}$$

en waarvoor:

$$\forall Q \subseteq A : \gamma^T(Q) \quad = \quad \mathrm{Nec}\left(p = T | Q = T\right) \tag{3.61}$$
$$\forall Q \subseteq A : \gamma^F(Q) \quad = \quad \mathrm{Nec}\left(p = F | Q = F\right). \tag{3.62}$$

Deze combinatie van deze twee vertrouwensmaten noemen we conditionele necessiteit. We leggen bij definitie vast dat de zekerheid over het waar (resp.

vals) zijn van p niet afhangt van actoren die postuleren dat p vals (resp. waar) is. We leggen met andere woorden de volgende voorwaarden op:

$$\forall Q \subseteq A : \text{Nec}\,(p = T | Q = T) = \text{Nec}\left(p = T | Q = T \wedge \overline{Q} = F\right) \quad (3.63)$$

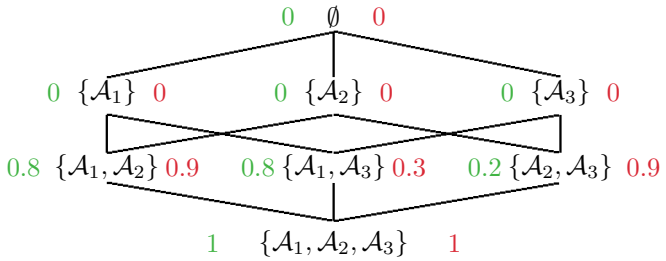
$$\forall Q \subseteq A : \text{Nec}\,(p = F | Q = F) = \text{Nec}\left(p = F | Q = F \wedge \overline{Q} = T\right) \quad (3.64)$$

$$\forall Q \subseteq A : \text{Nec}\,(p = T | Q = T) = \text{Nec}\left(p = T | \overline{Q} = F\right) \quad (3.65)$$

$$\forall Q \subseteq A : \text{Nec}\,(p = F | Q = F) = \text{Nec}\left(p = F | \overline{Q} = T\right). \quad (3.66)$$

Voor willekeurige $Q \subseteq A$ wordt door $\gamma^T(Q)$ (resp. $\gamma^F(Q)$) aangegeven wat de zekerheid is dat p waar (resp. vals) is, indien elke actor uit Q postuleert dat p waar (resp. vals) is. De vertrouwensmaten beschrijven bijgevolg de zekerheid over het waar (resp. vals) zijn van p, die kan worden afgeleid uit de postulaten van een deel van de actoren. Wanneer elke actor postuleert dat p waar is, dan moeten we besluiten dat p waar is. Wanneer elke actor postuleert dat p vals is, dan moeten we besluiten dat p vals is. Wanneer we geen enkel postulaat in beschouwing nemen, is de afgeleide necessiteit voor zowel waar als vals gelijk aan 0. Ten slotte kan de conditionele necessiteit over p niet dalen naarmate meer actoren dezelfde waarheidswaarde postuleren. Gelet op (1.49), (1.50) en (1.51), moeten beide functies γ^T en γ^F inderdaad noodzakelijk vertrouwensmaten zijn.

Voorbeeld 3.2
Beschouw een propositie p en beschouw een verzameling van drie actoren $A = \{\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3\}$ waarbij elke actor kennis uitdrukt over p. De machtsverzameling van A kan worden voorgesteld met een Hasse diagram. Figuur 3.3 toont een dergelijk Hasse diagram. Bij elke deelverzameling $Q \subseteq A$ staat links $\gamma^T(Q)$ en rechts $\gamma^F(Q)$.



Figuur 3.3: Voorbeeld van conditionele necessiteit

Uit Figuur 3.3 leiden we af dat het vertrouwen in een actor afzonderlijk steeds gelijk is aan 0. Wanneer $\mathcal{A}_1$ en $\mathcal{A}_2$ postuleren dat p waar is, dan hebben een vertrouwen van 0.8 in het waar zijn van p. Wanneer $\mathcal{A}_2$ en $\mathcal{A}_3$ postuleren dat p vals is, dan hebben we een vertrouwen van 0.9 in het vals zijn van p.

We merken op dat γ^T en γ^F voor elke $Q \subseteq A$ samen een necessiteitsverdeling beschrijven over $\mathbb{B}$. Bijgevolg moet de normalisatiewet gelden. Gelet op (3.63)

en (3.64) kunnen we de normalisatiewet van necesiteiten toepassen. Deze stelt dat:

$$\min\left(\mathrm{Nec}(S|C), \mathrm{Nec}(\overline{S}|C)\right) = 0. \tag{3.67}$$

Bijgevolg moet er gelden dat:

$$\forall Q \subseteq A : \min\left(\gamma^T(Q), \gamma^F(\overline{Q})\right) = 0. \tag{3.68}$$

Een gevolg van (3.68) is dat:

$$\forall Q \subseteq A : \gamma^T(Q) > 0 \quad \Rightarrow \quad \left(\forall Q' \subseteq A : Q \cap Q' = \emptyset \Rightarrow \gamma^F(Q') = 0\right) \tag{3.69}$$

$$\forall Q \subseteq A : \gamma^F(Q) > 0 \quad \Rightarrow \quad \left(\forall Q' \subseteq A : Q \cap Q' = \emptyset \Rightarrow \gamma^T(Q') = 0\right). \tag{3.70}$$

Hoewel het gebruik van twee vertrouwensmaten voor n actoren leidt tot de specificiatie van 2^{n+1} necessiteiten, kunnen we aantonen dat de toekenning van necessiteiten aan strenge regels onderhevig is. Meer bepaald geldt het volgende.

Stelling 3.2
Beschouw een propositie p en een verzameling $A = \{\mathcal{A}_1, ..., \mathcal{A}_n\}$ van actoren. Beschouw de vertrouwensmaten γ^T en γ^F en beschouw voor willekeurige $k \in \{1, ..., n-1\}$ een verzameling $V = \{Q_1, ..., Q_{k+1}\}$ waarbij elke Q_i een verzameling van actoren is zodat er geldt:

$$\forall Q \in V : |Q| = k \tag{3.71}$$

en

$$\forall Q \in V : \forall Q' \in V : (Q \neq Q') \Rightarrow (|Q \cap Q'| = 0). \tag{3.72}$$

Voor een dergelijke verzameling V geldt er dat:

$$\left(\forall Q \in V : \gamma^T(Q) \neq 0\right) \quad \Rightarrow \left(\forall Q \subseteq A \wedge |Q| = k : \gamma^F(Q) = 0\right) \tag{3.73}$$

$$\left(\forall Q \in V : \gamma^F(Q) \neq 0\right) \quad \Rightarrow \left(\forall Q \subseteq A \wedge |Q| = k : \gamma^T(Q) = 0\right). \tag{3.74}$$

Bewijs. We bewijzen de stelling in het geval van γ^T. Beschouw een deelverzameling $Q'' \subseteq A$ zodat:

$$\forall Q \in V : Q \not\subseteq Q'' \tag{3.75}$$

We kunnen nu berekenen hoeveel actoren Q'' maximaal kan bevatten. Aangezien V $k{+}1$ disjuncte verzamelingen met kardinaliteit k bevat, zijn er $n{-}k(k{+}1)$ actoren uit A die in geen enkele verzameling Q uit V voorkomen. Aangezien Q'' geen *superverzameling* mag zijn van een Q uit V, kunnen we uit elke $Q \in V$ maximaal $k-1$ elementen toevoegen aan Q''. Dit betekent dat Q'' maximaal:

$$n - (k\,(k+1)) + (k+1)\,(k-1) = n - k - 1 \tag{3.76}$$

actoren uit A kan bevatten. Bijgevolg geldt er dat:

$$\forall Q' \subseteq A \wedge |Q'| = n - k : \exists Q \in V : Q \subseteq Q'. \tag{3.77}$$

Door monotoniteit van γ^T geldt er dan:

$$\forall Q' \subseteq A \wedge |Q'| = n - k : \gamma^T(Q') \neq 0 \tag{3.78}$$

zodat er omwille van de normalisatiewet (3.68) geldt dat:

$$\forall Q \subseteq A \wedge |Q| = k : \gamma^F(Q) = 0. \tag{3.79}$$

□

De voorwaarde van Stelling 3.2 is nooit voldaan voor verzamelingen Q waarbij $|Q| > \lfloor n/2 \rfloor$. De stelling toont aan dat het opbouwen van de conditionele necessiteit vrij snel beperkt wordt, waardoor de constructie sterk wordt vereenvoudigd. Zo kan ook de volgende eigenschap worden aangetoond.

Eigenschap 3.1
Gegeven vertrouwensmaten γ^T en γ^F voor een verzameling A van n actoren. Er zijn minstens 2^n necessiteiten gelijk aan 0:

$$\left|\left\{Q|Q \subseteq A \wedge \left(\gamma^T(Q) = 0\right)\right\}\right| + \left|\left\{Q|Q \subseteq A \wedge \left(\gamma^F(Q) = 0\right)\right\}\right| \geq 2^n. \tag{3.80}$$

Bewijs. De eigenschap is een direct gevolg van de normalisatiewet (3.68). □

Een bijzonder geval van conditionele necessiteit is het geval van drastische vertrouwensmaten. De conditionele necessiteit is F-drastisch als:

$$\forall Q \subseteq A : \gamma^T(Q) = 0 \Rightarrow \gamma^F(\overline{Q}) = 1 \tag{3.81}$$

en T-drastisch als:

$$\forall Q \subseteq A : \gamma^F(Q) = 0 \Rightarrow \gamma^T(\overline{Q}) = 1. \tag{3.82}$$

In het geval van drastische vertrouwensmaten geldt de volgende stelling.

Stelling 3.3
Beschouw een drastische vertrouwensmaat γ^F. Indien γ^T een possibiliteitsmaat is, dan is γ^F een necessiteitsmaat. Indien γ^T een necessiteitsmaat is, dan is γ^F een possibiliteitsmaat.

Bewijs. Stel dat γ^T een possibiliteitsmaat is. Enerzijds hebben we dat:

$$\left(\gamma^F(Q_1 \cap Q_2) = 1\right) \Leftrightarrow \left(\gamma^T(\overline{Q_1} \cup \overline{Q_2}) = 0 = \max\left(\gamma^T(\overline{Q_1}), \gamma^T(\overline{Q_2})\right)\right). \tag{3.83}$$

Bijgevolg geldt er ook dat:

$$\gamma^F(Q_1) = 1 \tag{3.84}$$

en dat:

$$\gamma^F(Q_2) = 1. \tag{3.85}$$

Anderzijds geldt er dat:

$$\left(\gamma^F(Q_1 \cap Q_2) = 0\right) \Leftrightarrow \left(\gamma^T(\overline{Q_1} \cup \overline{Q_2}) = 1 = \max\left(\gamma^T(\overline{Q_1}), \gamma^T(\overline{Q_2})\right)\right) \tag{3.86}$$

wat betekent dat:

$$\left(\gamma^F(Q_1) = 0\right) \vee \left(\gamma^F(Q_2) = 0\right). \tag{3.87}$$

Uit beiden volgt nu dat:

$$\gamma^F(Q_1 \cap Q_2) = \min\left(\gamma^F(Q_1), \gamma^F(Q_2)\right). \tag{3.88}$$

Het geval waar γ^T een necessiteitsmaat is, laat zich volledig analoog bewijzen. □

In Hoofdstuk 7 zullen we verder bestuderen hoe de conditionele necessiteit kan worden bepaald in de context van complexe objecten.

3.3.4 Combinatiefunctie

We beschikken nu over het concept van conditionele necessiteit dat toelaat om het vertrouwen in actoren uit te drukken. We hebben aangetoond hoe dit vertrouwen voortkomt uit waarheidsbehoudende afhankelijkheden tussen actoren. In een volgende stap bestuderen we hoe dit alles leidt tot een voorstelling van de kennis over de waarheidswaarde van een propositie p. Om dit te bewerkstelligen, is het nodig om marginale kennis in de vorm van een possibilistische waarheidswaarde en conditionele kennis te combineren met elkaar. Hiervoor kunnen we de conditioneringswet voor necessiteiten gebruiken (Hoofdstuk 1). Deze wet schrijft voor dat als een actor $\mathcal{A} \in A$ kennis over de waarheidswaarde van een propositie p geeft als $\widetilde{p}$, dan kunnen we deze kennis combineren met conditionele kennis als volgt:

$$\begin{aligned} &\mathrm{Nec}\left(\mathcal{A} \rightsquigarrow (p = T)\right) \quad \wedge \quad \mathrm{Nec}\left(p = T | \mathcal{A} \rightsquigarrow (p = T)\right) && (3.89)\\ &\mathrm{Nec}\left(\mathcal{A} \rightsquigarrow (p = F)\right) \quad \wedge \quad \mathrm{Nec}\left(p = F | \mathcal{A} \rightsquigarrow (p = F)\right). && (3.90) \end{aligned}$$

Door gebruik te maken van eerder ingevoerde notaties kunnen we dit herschrijven als:

$$\begin{aligned} &\left(1 - \mu_{\widetilde{p}}(F)\right) \quad \wedge \quad \gamma^T\left(\{\mathcal{A}\}\right) && (3.91)\\ &\left(1 - \mu_{\widetilde{p}}(T)\right) \quad \wedge \quad \gamma^F\left(\{\mathcal{A}\}\right). && (3.92) \end{aligned}$$

Deze uitdrukkingen geven de correcte combinatie voor postulaten van afzonderlijke actoren. We moeten dergelijke combinaties echter ook kunnen maken voor willekeurige deelverzamelingen van A. Meer bepaald hebben we combinaties van de volgende vorm nodig:

$$\begin{aligned} &\mathrm{Nec}(Q = T) \quad \wedge \quad \gamma^T(Q) && (3.93)\\ &\mathrm{Nec}(Q = F) \quad \wedge \quad \gamma^F(Q). && (3.94) \end{aligned}$$

Gelet op de eigenschappen van necessiteit kunnen we dit vereenvoudigen aangezien er geldt dat:

$$\begin{aligned} \mathrm{Nec}(Q = T) &= \min_{\mathcal{A}_i \in Q}\left(1 - \mu_{\widetilde{p}_i}(F)\right) && (3.95)\\ \mathrm{Nec}(Q = F) &= \min_{\mathcal{A}_i \in Q}\left(1 - \mu_{\widetilde{p}_i}(T)\right). && (3.96) \end{aligned}$$

waarbij $\widetilde{p}_i$ de kennis voorstelt die wordt gegeven door $\mathcal{A}_i$. Met de combinatieregel van marginale en conditionele necessiteit voorhanden, blijft de vraag hoe een possibilistische waarheidswaarde kan worden afgeleid, zodat die de kennis over het waar en vals zijn van p beschrijft.

Wanneer we n actoren beschouwen, bestaan er 2^n verschillende combinaties van postulaten van de actoren. Elke $Q \subseteq A$ komt overeen met een combinatie van postulaten, d.i. de actoren in Q stellen dat p waar is en de actoren in $\overline{Q}$ stellen dat p vals is. Enerzijds kunnen we voor elke $Q \subseteq A$ aftoetsen wat de zekerheid is over de postulaten $Q = T$ en $Q = F$ door gebruik te maken van Q_π. Anderzijds bepalen $\gamma^T(Q)$ en $\gamma^F(Q)$ de zekerheid dat p waar (resp. vals) is als $Q = T$ (resp. $Q = F$). Anders gezegd, voor een willekeurige combinatie van postulaten kunnen we de zekerheid van deze postulaten berekenen op basis van Q_π en kunnen we, gesteld dat de postulaten gelden, de zekerheid over de waarheidswaarde van p berekenen op basis van γ^T en γ^F. Dit betekent dat we zoeken naar die $Q \subseteq A$ waarvoor $Q = T$ (resp. $Q = F$) maximaal wordt bevestigd door Q_π én waarvoor de afgeleide zekerheid voor p maximaal is. De af te leiden necessiteiten voor waar en vals moeten hierdoor het gevolg zijn van een maximalisatie van (3.93) en (3.94) over Q. Dit wil zeggen dat:

$$\mathrm{Nec}(p = T) = \bigvee_{Q \subseteq A} \left(\mathrm{Nec}\left(Q = T\right) \wedge \gamma^T\left(Q\right)\right) \tag{3.97}$$

$$\mathrm{Nec}(p = F) = \bigvee_{Q \subseteq A} \left(\mathrm{Nec}\left(Q = F\right) \wedge \gamma^F\left(Q\right)\right). \tag{3.98}$$

Merk op dat deze afgeleide formules voor $\mathrm{Nec}(p = T)$ en $\mathrm{Nec}(p = F)$ beiden de vorm van een discrete Sugeno integraal vertonen [61]. Laat ons nu eerst aantonen dat deze formules vereenvoudigd kunnen worden. Laat $.^{()^T}$ een permutatie op de elementen van P_π zijn zodat $\forall i \in \{1, ..., n-1\} : \widetilde{p}_{(i)^T} \geq \widetilde{p}_{(i+1)^T}$. Laat $.^{()^F}$ een permutatie op de elementen van P_π zijn zodat $\forall i \in \{1, ..., n-1\} : \widetilde{p}_{(i)^F} \leq \widetilde{p}_{(i+1)^F}$. Laat $A_{(i)^T}$ de verzameling $\{\mathcal{A}_{(1)^T}, ..., \mathcal{A}_{(i)^T}\}$ voorstellen en laat $A_{(i)^F}$ de verzameling $\{\mathcal{A}_{(1)^F}, ..., \mathcal{A}_{(i)^F}\}$ zijn waarbij $\mathcal{A}_{(i)^T}$ de actor is die $\widetilde{p}_{(i)^T}$ genereert en waarbij $\mathcal{A}_{(i)^F}$ de actor is die overeenkomt met $\widetilde{p}_{(i)^F}$. We kunnen nu de volgende stelling bewijzen.

Stelling 3.4
Gegeven een propositie p en een verzameling $A = \{\mathcal{A}_1, ..., \mathcal{A}_n\}$ van actoren. De necessiteiten voor de waarheidswaarde van p zijn gelijk aan:

$$\bigvee_{Q \subseteq A} \left(\mathrm{Nec}\left(Q = T\right) \wedge \gamma^T\left(Q\right)\right) = \bigvee_{i=1}^{n} \left(\mathrm{Nec}\left(\mathcal{A}_{(i)^T} = T\right) \wedge \gamma^T\left(A_{(i)^T}\right)\right) \tag{3.99}$$

$$\bigvee_{Q \subseteq A} \left(\mathrm{Nec}\left(Q = F\right) \wedge \gamma^F\left(Q\right)\right) = \bigvee_{i=1}^{n} \left(\mathrm{Nec}\left(\mathcal{A}_{(i)^F} = F\right) \wedge \gamma^F\left(A_{(i)^F}\right)\right). \tag{3.100}$$

Bewijs. We leveren het bewijs voor het geval van waar. Het geval voor

vals is volledig analoog te bewijzen. Beschouw een verzameling $Q \subseteq A$ zodat:

$$\exists \mathcal{A}_i \in A \backslash Q : \widetilde{p}_i \geq \widetilde{\bigwedge_{\widetilde{q} \in Q_\pi}} (\widetilde{q})$$

dan geldt er enerzijds:

$$\gamma^T (Q \cup \{\mathcal{A}_i\}) \geq \gamma^T (Q) \tag{3.101}$$

en anderzijds:

$$\text{Nec} (Q \cup \{\mathcal{A}_i\} = T) \geq \text{Nec} (Q = T) \tag{3.102}$$

waaruit volgt dat:

$$\gamma^T (Q \cup \{\mathcal{A}_i\}) \wedge \text{Nec} (Q \cup \{\mathcal{A}_i\} = T) \geq \gamma^T (Q) \wedge \text{Nec} (Q = T). \tag{3.103}$$

Hieruit besluiten we dat een groep van actoren Q met kardinaliteit k slechts kan leiden tot de gezochte necessiteit als en alleen als Q_π de k grootste possibilistische waarheidswaarden bevat. Voor een dergelijke verzameling Q wordt $\text{Nec}(Q = T)$ berekend als de necessiteit voor waar van de k^{de} possibilistische waarheidswaarde uit P_π onder de ordening $._{()^T}$. □

Omwille van de gelijkenis met de klassieke discrete Sugeno-integraal [61], definiëren we een functie die we de Sugeno-integraal voor possibilistische waarheidswaarden noemen.

Definitie 3.14 (Sugeno-integraal)
Gegeven een propositie p en een verzameling van n actoren $A = \{\mathcal{A}_1, ..., \mathcal{A}_n\}$. Laat P_π de multiverzameling van possibilistische waarheidswaarden gegenereerd door de actoren in A voorstellen. De Sugeno-integraal voor possibilistische waarheidswaarden is een functie:

$$S_{\gamma^{T,F}} (P_\pi) : \mathcal{M}_n (\mathcal{F}(\mathbb{B})) \to \mathcal{F}(\mathbb{B}) \tag{3.104}$$

waarbij $S_{\gamma^{T,F}} (P_\pi) = \widetilde{p}$ zodat:

$$\text{Nec}(p = T) = \bigvee_{i=1}^{n} \left(\text{Nec} \left(\mathcal{A}_{(i)^T} = T\right) \wedge \gamma^T \left(A_{(i)^T}\right)\right) \tag{3.105}$$

$$\text{Nec}(p = F) = \bigvee_{i=1}^{n} \left(\text{Nec} \left(\mathcal{A}_{(i)^F} = F\right) \wedge \gamma^F \left(A_{(i)^F}\right)\right). \tag{3.106}$$

Voorbeeld 3.3
Beschouw de propositie p en de conditionele necessiteit uit Voorbeeld 3.2. Stel dat de kennis over de waarheidswaarde van p gegeven door de drie actoren gelijk is aan:

$$\widetilde{p}_1 = (1, 0.1) \tag{3.107}$$

$$\widetilde{p}_2 = (0.3, 1) \tag{3.108}$$

$$\widetilde{p}_3 = (1, 0.4). \tag{3.109}$$

dan kunnen we deze sorteren op meeste zekerheid voor waar eerst:

$$\widetilde{p}_{(1)^T} = \widetilde{p}_1 \tag{3.110}$$
$$\widetilde{p}_{(2)^T} = \widetilde{p}_3 \tag{3.111}$$
$$\widetilde{p}_{(3)^T} = \widetilde{p}_2. \tag{3.112}$$

De omgekeerde sortering is dan:

$$\widetilde{p}_{(1)^F} = \widetilde{p}_2 \tag{3.113}$$
$$\widetilde{p}_{(2)^F} = \widetilde{p}_3 \tag{3.114}$$
$$\widetilde{p}_{(3)^F} = \widetilde{p}_1. \tag{3.115}$$

Voor de uitwerking van de Sugeno-integraal voor possibilistische waarheidswaarden vinden we, gelet op:

$$A_{(1)^T} = \{\mathcal{A}_1\} \tag{3.116}$$
$$A_{(2)^T} = \{\mathcal{A}_1, \mathcal{A}_3\} \tag{3.117}$$
$$A_{(3)^T} = \{\mathcal{A}_1, \mathcal{A}_3, \mathcal{A}_2\} \tag{3.118}$$

enerzijds:

$$\mathrm{Nec}(p = T) = \bigvee_{i=1}^{3} \left(\mathrm{Nec}\left(\mathcal{A}_{(i)^T} = T\right) \wedge \gamma^T\left(A_{(i)^T}\right)\right) \tag{3.119}$$
$$= \max(\min(0.9, 0), \min(0.6, 0.8), \min(0, 1)) \tag{3.120}$$
$$= 0.6 \tag{3.121}$$

en gelet op:

$$A_{(1)^F} = \{\mathcal{A}_2\} \tag{3.122}$$
$$A_{(2)^F} = \{\mathcal{A}_2, \mathcal{A}_3\} \tag{3.123}$$
$$A_{(3)^F} = \{\mathcal{A}_2, \mathcal{A}_3, \mathcal{A}_1\} \tag{3.124}$$

anderzijds:

$$\mathrm{Nec}(p = F) = \bigvee_{i=1}^{3} \left(\mathrm{Nec}\left(\mathcal{A}_{(i)^F} = F\right) \wedge \gamma^F\left(A_{(i)^F}\right)\right) \tag{3.125}$$
$$= \max(\min(0.7, 0), \min(0, 0.9), \min(0, 1)) \tag{3.126}$$
$$= 0.0. \tag{3.127}$$

Hieruit volgt dat:

$$S_{\gamma^{T,F}}(\{\widetilde{p}_1, \widetilde{p}_2, \widetilde{p}_3\}) = (1 - 0, 1 - 0.6) = (1, 0.4)\,. \tag{3.128}$$

De functie $S_{\gamma^{T,F}}$ uit Definitie 3.14 geeft een possibilistische waarheidswaarde die de kennis weerspiegeld over de waarheidswaarde van een propositie, rekening houdend met het vertrouwen dat in actoren wordt gesteld. We kunnen

deze possibilistische waarheidswaarde ook berekenen door een redenering met possibiliteiten in plaats van necessiteiten. Dit wordt aangetoond door de volgende stelling.

Stelling 3.5
Gegeven een Boolese propositie, een verzameling A van n actoren en conditionele necessiteit γ^T en γ^F. Indien:

$$S_{\gamma^{T,F}}(P_\pi) = \widetilde{p} \tag{3.129}$$

dan geldt er:

$$\text{Pos}(p = T) = \bigvee_{i=1}^{n} \left(\text{Pos}\left(\mathcal{A}_{(i)^T} = T\right) \wedge \text{Pos}\left(p = T | A_{(i)^T} = T\right)\right) \tag{3.130}$$

$$\text{Pos}(p = F) = \bigvee_{i=1}^{n} \left(\text{Pos}\left(\mathcal{A}_{(i)^F} = F\right) \wedge \text{Pos}\left(p = F | A_{(i)^F} = F\right)\right) \tag{3.131}$$

Stelling 3.5 stelt dat de Sugeno-integraal equivalent kan worden genoteerd met possibiliteiten. Dit kunnen we bewijzen door gebruik te maken van de volgende stelling.

Stelling 3.6
Veronderstel een vector $\mathbf{x} \in [0,1]^n$ zodat $\mathbf{x}_i \geq \mathbf{x}_{i+1}$ en een vector $\mathbf{y} \in [0,1]^n$ zodat $\mathbf{y}_i \leq \mathbf{y}_{i+1}$ en $\mathbf{y}_n = 1$. Beschouw vervolgens de vectoren:

$$\mathbf{x}' = (1 - \mathbf{x}_1, ..., 1 - \mathbf{x}_n) \tag{3.132}$$

$$\mathbf{y}' = (1, 1 - \mathbf{y}_1, ..., 1 - \mathbf{y}_{n-1}). \tag{3.133}$$

Er geldt dan:

$$\max_{i \in \{1,...,n\}} \min(\mathbf{x}_i, \mathbf{y}_i) + \max_{i \in \{1,...,n\}} \min(\mathbf{x}'_i, \mathbf{y}'_i) = 1. \tag{3.134}$$

Bewijs Stelling 3.6. Beschouwen we de variabele z zodat

$$z = \max_{i \in \{1,...,n\}} \min(\mathbf{x}_i, \mathbf{y}_i). \tag{3.135}$$

Er moet dan gelden dat:

$$z \in \mathbf{x} \vee z \in \mathbf{y}. \tag{3.136}$$

(a) Veronderstel $z \in \mathbf{x}$ dan hebben we dat:

$$\exists l \in \{1, ..., n\} : z = \mathbf{x}_l = \min(\mathbf{x}_l, \mathbf{y}_l) \tag{3.137}$$

(*) In het grensgeval waar $l = n$, gelet op $\forall k < n : \mathbf{x}_k \geq \mathbf{x}_n$, kan worden ingezien dat:

$$\forall k < n : \mathbf{y}_k \leq \mathbf{x}_n. \tag{3.138}$$

Zoniet, dan bestaat er een $\mathbf{y}_k > \mathbf{x}_n$ zodat z niet gelijk zou zijn aan $\mathbf{x}_n$. We kunnen nu zien dat (3.138) equivalent is met:

$$\forall k < n : 1 - \mathbf{y}_k \geq 1 - \mathbf{x}_n. \tag{3.139}$$

Bovendien is $\mathbf{y}'_1 = 1$ en dus geldt er:

$$\forall k \leq n : \mathbf{y}'_i \geq \mathbf{x}'_n. \tag{3.140}$$

Dit betekent dat:

$$\max_{i\in\{1,\ldots,n\}} \min(\mathbf{x}_i, \mathbf{y}_i) + \max_{i\in\{1,\ldots,n\}} \min(\mathbf{x}'_i, \mathbf{y}'_i) = \mathbf{x}_n + 1 - \mathbf{x}_n = 1. \tag{3.141}$$

(*) Indien $l < n$ is $\mathbf{y}_l \geq \mathbf{x}_l$ en bijgevolg:

$$(1 - \mathbf{y}_l \leq 1 - \mathbf{x}_l) \Leftrightarrow \left(\mathbf{y}'_{l+1} \leq \mathbf{x}'_l\right). \tag{3.142}$$

Aangezien $\mathbf{y}'$ een dalende vector is, volgt er:

$$\forall k \geq l + 1 : \mathbf{y}'_k \leq \mathbf{x}'_l \Rightarrow \forall k > l : \min(\mathbf{x}'_k, \mathbf{y}'_k) \leq \mathbf{x}'_l. \tag{3.143}$$

Ook geldt er dat $\mathbf{x}'$ een stijgende vector is, waaruit volgt dat:

$$\forall k < l : \mathbf{x}'_k \leq \mathbf{x}'_l \Rightarrow \forall k < l : \min(\mathbf{x}'_k, \mathbf{y}'_k) \leq \mathbf{x}'_l. \tag{3.144}$$

Vervolgens is het ook zo dat

$$\forall k < l : (\mathbf{x}_k \geq \mathbf{x}_l) \wedge (\min(\mathbf{x}_k, \mathbf{y}_k) \leq \mathbf{x}_l) \tag{3.145}$$

waaruit volgt dat

$$\forall k < l : \mathbf{y}_k \leq \mathbf{x}_l. \tag{3.146}$$

Uit dit resultaat kan worden afgeleid dat als $l > 1$ er geldt dat:

$$\min(\mathbf{x}'_l, \mathbf{y}'_l) = 1 - \max(\mathbf{x}_l, \mathbf{y}_{l-1}) = 1 - \mathbf{x}_l = \mathbf{x}'_l. \tag{3.147}$$

Indien $l = 1$ dan is:

$$\min(\mathbf{x}'_l, \mathbf{y}'_l) = \min(\mathbf{x}'_l, 1) = \mathbf{x}'_l. \tag{3.148}$$

Bijgevolg is:

$$\max_{i=1,\ldots,n} \min(\mathbf{x}'_i, \mathbf{y}'_i) = \mathbf{x}'_l \tag{3.149}$$

waaruit we besluiten dat:

$$\max_{i=1,\ldots,n} \min(\mathbf{x}_i, \mathbf{y}_i) + \max_{i=1,\ldots,n} \min(\mathbf{x}'_i, \mathbf{y}'_i) = \mathbf{x}_l + \mathbf{x}'_l = \mathbf{x}_l + 1 - \mathbf{x}_l = 1. \tag{3.150}$$

(b) Veronderstel dat $z \in \mathbf{y}$ dan kunnen we een gelijkaardige redenering volgen:

$$\exists l \in \{1, ..., n\} : z = \mathbf{y}_l = \min(\mathbf{x}_l, \mathbf{y}_l) \tag{3.151}$$

(*) In het grensgeval waar $l = n$, gelet op enerzijds het feit dat $\mathbf{x}$ dalend is en anderzijds het feit dat $\mathbf{y}_n = 1$, volgt er dat:

$$\forall k \leq n : \mathbf{x}_k = 1 \tag{3.152}$$

en dus

$$\forall k \leq n : \mathbf{x}'_k = 0. \tag{3.153}$$

Bijgevolg geldt er:

$$\max_{i\in\{1,\ldots,n\}} \min(\mathbf{x}_i, \mathbf{y}_i) + \max_{i\in\{1,\ldots,n\}} \min(\mathbf{x}'_i, \mathbf{y}'_i) = 1 + 0 = 1. \tag{3.154}$$

(*) Indien $l < n$, dan is $\mathbf{y}_l \leq \mathbf{x}_l$ en bijgevolg:

$$(1 - \mathbf{y}_l \geq 1 - \mathbf{x}_l) \Leftrightarrow \left(\mathbf{y}'_{l+1} \geq \mathbf{x}'_l\right). \tag{3.155}$$

Omdat $\mathbf{x}'$ een stijgende vector is, geldt er dat:

$$\forall k < l + 1 : \mathbf{x}'_k \leq \mathbf{y}'_{l+1} \Rightarrow \forall k < l + 1 : \min(\mathbf{x}'_k, \mathbf{y}'_k) \leq \mathbf{y}'_{l+1}. \tag{3.156}$$

Aangezien $\mathbf{y}'$ een dalende vector is, geldt er dat:

$$\forall k > l + 1 : \mathbf{y}'_k \leq \mathbf{y}'_{l+1} \Rightarrow \forall k > l + 1 : \min(\mathbf{x}'_k, \mathbf{y}'_k) \leq \mathbf{y}'_{l+1}. \tag{3.157}$$

Vervolgens is het ook zo dat:

$$\forall k > l : (\mathbf{y}_k \geq \mathbf{y}_l) \wedge (\min(\mathbf{x}_k, \mathbf{y}_k) \leq \mathbf{y}_l) \tag{3.158}$$

waaruit kan worden afgeleid dat:

$$\forall k > l : \mathbf{x}_k \leq \mathbf{y}_l. \tag{3.159}$$

Hieruit volgt er dat:

$$\min(\mathbf{x}'_{l+1}, \mathbf{y}'_{l+1}) = 1 - \max(\mathbf{x}_{l+1}, \mathbf{y}_l) = 1 - \mathbf{y}_l = \mathbf{y}'_{l+1}. \tag{3.160}$$

Bijgevolg is:

$$\max_{i=1,\ldots,n} \min(\mathbf{x}'_i, \mathbf{y}'_i) = \mathbf{y}'_{l+1} \tag{3.161}$$

en krijgen we opnieuw:

$$\max_{i=1,\ldots,n} \min(\mathbf{x}_i, \mathbf{y}_i) + \max_{i=1,\ldots,n} \min(\mathbf{x}'_i, \mathbf{y}'_i) = \mathbf{y}_l + \mathbf{y}'_{l+1} = \mathbf{y}_l + 1 - \mathbf{y}_l = 1. \tag{3.162}$$

De combinatie van gevallen (a) en (b) bewijst het gestelde. □

Bewijs Stelling 3.5. De stelling volgt uit het herschrijven van de uitdrukkingen voor possibiliteit en necessiteit, zodat Stelling 3.6 kan worden toegepast. Voor het herschrijven van de uitdrukking voor de possibiliteit van waar, stellen we:

$$k = |\{\widetilde{p}_i | \widetilde{p}_i \in P_\pi \wedge (\mathrm{Nec}(\mathcal{A}_i = F) = 0)\}| \tag{3.163}$$

Vermits de vertrouwensmaat Pos(.|.) in (3.130) stijgend is, kunnen we schrijven dat:

$$\begin{aligned} \text{Pos}(p=T) &= \bigvee_{i=1}^{n} \left(\text{Pos}\left(\mathcal{A}_{(i)^T}=T\right) \wedge \text{Pos}\left(p=T|A_{(i)^T}=T\right)\right) && (3.164)\\ &= \bigvee_{i=k}^{n} \left(\text{Pos}\left(\mathcal{A}_{(i)^T}=T\right) \wedge \text{Pos}\left(p=T|A_{(i)^T}=T\right)\right) && (3.165) \end{aligned}$$

Dit kunnen we herschrijven in termen van necessiteiten:

$$\text{Pos}(p=T) = \bigvee_{i=k}^{n} \left(1-\text{Nec}\left(\mathcal{A}_{(i)^T}=F\right) \wedge 1-\text{Nec}\left(p=F|A_{(i)^T}=T\right)\right). \tag{3.166}$$

Aangezien de permutaties $.{}_{(i)^T}$ en $.{}_{(i)^F}$ aan elkaar zijn gekoppeld via de uitdrukking:

$$\widetilde{p}_{(i)^T} = \widetilde{p}_{(n-i+1)^F} \tag{3.167}$$

en gelet op (3.64) kunnen we schrijven dat:

$$\text{Pos}(p=T) = \bigvee_{i=k}^{n} \left(1-\text{Nec}\left(\mathcal{A}_{(n-i+1)^F}=F\right) \wedge 1-\text{Nec}\left(p=F|A_{(n-i)^F}=F\right)\right). \tag{3.168}$$

Uitvoeren van de variabeletransformatie $j=n-i$ levert:

$$\text{Pos}(p=T) = \bigvee_{j=n-k}^{0} \left(1-\text{Nec}\left(\mathcal{A}_{(j+1)^F}=F\right) \wedge 1-\gamma^F\left(A_{(j)^F}\right)\right). \tag{3.169}$$

Anderzijds kunnen we de uitdrukking voor de necessiteit van vals schrijven als:

$$\text{Nec}(p=F) = \bigvee_{i=1}^{n} \left(\text{Nec}\left(\mathcal{A}_{(i)^F}=F\right) \wedge \gamma^F\left(A_{(i)^F}\right)\right) \tag{3.170}$$

Gelet op (3.163) en met de conventie dat $\text{Nec}(\mathcal{A}_{(0)^F}=F)=1$ volgt er dat:

$$\text{Nec}(p=F) = \bigvee_{i=0}^{n-k} \left(\text{Nec}\left(\mathcal{A}_{(i)^F}=F\right) \wedge \gamma^F\left(A_{(i)^F}\right)\right). \tag{3.171}$$

Gelet op (3.169) en (3.171) kunnen we de volgende vectoren opstellen:

$$\begin{aligned} \mathbf{x} &= \left(\gamma^F(A_{(n-k)^F}), ..., \gamma^F(A_{(0)^F})\right) && (3.172)\\ \mathbf{y} &= \left(\text{Nec}(A_{(n-k)^F}=F), ..., \text{Nec}(A_{(1)^F}=F), 1\right) && (3.173)\\ \mathbf{x}' &= \left(1-\gamma^F(A_{(n-k)^F}), ..., 1-\gamma^F(A_{(0)^F})\right) && (3.174)\\ \mathbf{y}' &= \left(1, 1-\text{Nec}(A_{(n-k)^F}=F), ..., 1-\text{Nec}(A_{(1)^F}=F)\right) && (3.175) \end{aligned}$$

Toepassing van Stelling 3.6 levert dan:

$$\text{Nec}(p = F) = 1 - \text{Pos}(p = T). \tag{3.176}$$

Op eenzelfde wijze kunnen bewijzen dat:

$$\text{Nec}(p = T) = 1 - \text{Pos}(p = F). \tag{3.177}$$

□

Hoewel we expliciet vermeld hebben dat we geen verband handhaven tussen de Sugeno-integraal in het raamwerk van possibilistische waarheidswaarden en logische functies in het Boolese domein, weerspiegelt de Sugeno-integraal in sommige gevallen het gedrag van een Boolese functie. Veronderstellen we dat de conditionele necessiteit F-drastisch en T-drastisch is en dat er geldt:

$$\forall \widetilde{p} \in P_\pi : \widetilde{p} = (1,0) \vee \widetilde{p} = (0,1) \tag{3.178}$$

dan kunnen we twee bijzondere gevallen onderscheiden. In het eerste geval is de conditionele necessiteit geconstrueerd als volgt:

$$\forall Q \subset A \quad : \quad \gamma^T(Q) = 0 \tag{3.179}$$
$$\forall Q \subseteq A \quad : \quad (Q \neq \emptyset) \Rightarrow \gamma^F(Q) = 1. \tag{3.180}$$

In dit geval drukken we uit dat het waar zijn van p enkel volgt wanneer alle actoren met zekerheid stellen dat p waar is. Dit geval komt overeen met een Boolese conjunctie van de postulaten van de actoren. Het kan makkelijk worden geverifieerd dat de functie $S_{\gamma^{T,F}}$ inderdaad dit gedrag vertoont zoals weergegeven in Tabel 3.1.

	$\widetilde{p}_1 = (1,0)$	$\widetilde{p}_1 = (0,1)$
$\widetilde{p}_2 = (1,0)$	$(1,0)$	$(0,1)$
$\widetilde{p}_2 = (0,1)$	$(0,1)$	$(0,1)$

Tabel 3.1: Conjunctief gedrag van $S_{\gamma^{T,F}}$ voor $|A| = 2$ zonder onzekerheid

In het tweede geval is de conditionele necessiteit geconstrueerd als volgt:

$$\forall Q \subset A \quad : \quad \gamma^F(Q) = 0 \tag{3.181}$$
$$\forall Q \subseteq A \quad : \quad (Q \neq \emptyset) \Rightarrow \gamma^T(Q) = 1. \tag{3.182}$$

In dit geval drukken we uit dat het waar zijn van p volgt van zodra één van de actoren met zekerheid stelt dat p waar is. Dit geval komt overeen met een Boolese disjunctie van de postulaten van de actoren. Het kan opnieuw makkelijk worden geverifieerd dat de functie $S_{\gamma^{T,F}}$ inderdaad dit gedrag vertoont zoals weergegeven in Tabel 3.2.

De Boolese conjunctie en disjunctie komen voort uit randgevallen van drastische vertrouwensmaten. Voor tussenliggende gevallen van drastische vertrouwensmaten krijgen we een welgekende familie van Boolese functies. Drastische

	$\widetilde{p}_1 = (1,0)$	$\widetilde{p}_1 = (0,1)$
$\widetilde{p}_2 = (1,0)$	$(1,0)$	$(1,0)$
$\widetilde{p}_2 = (0,1)$	$(1,0)$	$(0,1)$

Tabel 3.2: Disjunctief gedrag van $S_{\gamma^{T,F}}$ voor $|A| = 2$ zonder onzekerheid

vertrouwensmaten komen steeds overeen met een functie van de vorm:

$$\bigwedge_{i=1}^{n} \left(q_{(i)^T} \Rightarrow \mathcal{A}_{(i)^T}\right) \tag{3.183}$$

waarbij:

$$\left(q_{(i)^T} = T\right) \quad \Leftrightarrow \quad \left(\gamma^T(A_{(i)^T}) = 1\right) \tag{3.184}$$

$$\left(q_{(i)^T} = F\right) \quad \Leftrightarrow \quad \left(\gamma^F(\overline{A_{(i)^T}}) = 1\right). \tag{3.185}$$

De propositie $q_{(i)}$ vormt in dit geval de voorwaarde waaraan voldaan moet zijn, opdat Q actoren bevat die voldoende vertrouwen hebben.

3.3.5 Eigenschappen van kenniscombinatie

In deze sectie onderzoeken we de eigenschappen van de functie $S_{\gamma^{T,F}}$. Een aantal van deze eigenschappen zijn eerder triviaal en zullen we niet bewijzen. Het kan bijvoorbeeld makkelijk worden ingezien dat $S_{\gamma^{T,F}}$ steeds commutatief is en puntsgewijs monotoon met betrekking tot P_π. Een belangrijke eigenschap die de nodige aandacht verdient, is vervat in de volgende stelling.

Stelling 3.7 (Normalisatie van $S_{\gamma^{T,F}}$)
Het resultaat van $S_{\gamma^{T,F}}$ is een genormaliseerde possibiliteitsverdeling over $\mathbb{B}$.

Bewijs. De necessiteit van $p = T$ wordt bereikt voor een welbepaalde index k tussen 1 en n zodat:

$$\exists k \in \{1, ..., n\} : \text{Nec}(p = T) = \text{Nec}(\mathcal{A}_{(k)^T} = T) \wedge \gamma^T(A_{(k)^T}). \tag{3.186}$$

Een dergelijke index l bestaat ook voor de necessiteit van $p = F$. We zullen echter deze index veronderstellen volgens de ordening $._{(i)^T}$ zodat:

$$\exists l \in \{1, ..., n\} : \text{Nec}(p = F) = \text{Nec}(\mathcal{A}_{(l)^T} = F) \wedge \gamma^F(\overline{A_{(l)^T}}) \tag{3.187}$$

(*) Als $\text{Nec}(p = T) = 0$, dan is het resultaat genormaliseerd.
(*) Als $\text{Nec}(p = T) > 0$, dan kunnen zich twee gevallen voordoen.
(a) Indien $l \leq k$ dan weten we, gelet op:

$$\forall i \in \{1, ..., k\} : \text{Nec}(\mathcal{A}_{(i)^T} = T) > 0 \tag{3.188}$$

en gelet op normalisatie van oorspronkelijke possibilistische waarheidswaarden dat:

$$\forall i \in \{1, ..., k\} : \text{Nec}(\mathcal{A}_{(i)^T} = F) = 0 \tag{3.189}$$

hetgeen betekent dat in dat geval $\text{Nec}(p = F) = 0$.

(b) Indien $l > k$ dan weten we, gelet op de monotoniteit van γ^T dat

$$\forall i \in \{k+1, ..., n\} : \gamma^T(A_{(i)^T}) > 0 \tag{3.190}$$

waaruit volgt dat:

$$\forall i \in \{k+1, ..., n\} : \gamma^F\left(\overline{A_{(i)^T}}\right) = 0 \tag{3.191}$$

hetgeen betekent dat ook in dit geval $\text{Nec}(p = F) = 0$. Bijgevolg vinden we dat:

$$\text{Nec}(p = T) > 0 \Rightarrow \text{Nec}(p = F) = 0. \tag{3.192}$$

Gelet op het feit dat door een volledige analoge redenering kan worden afgeleid dat:

$$\text{Nec}(p = F) > 0 \Rightarrow \text{Nec}(p = T) = 0 \tag{3.193}$$

is daarmee bewezen dat het resultaat van $S_{\gamma^{T,F}}$ steeds genormaliseerd is. □

Stelling 3.8
De functie $S_{\gamma^{T,F}}$ is ingeklemd tussen de Zadeh-uitbreiding van $\wedge$ en $\vee$. Dit betekent dat er voor een willekeurige collectie van possibilistische waarheidswaarden $P_\pi = \{\widetilde{p}_1, ..., \widetilde{p}_n\}$ geldt:

$$\left(\widetilde{p}_1 \,\widetilde{\wedge}\, ... \,\widetilde{\wedge}\, \widetilde{p}_n\right) \leq S_{\gamma^{T,F}}(P_\pi) \leq \left(\widetilde{p}_1 \,\widetilde{\vee}\, ... \,\widetilde{\vee}\, \widetilde{p}_n\right). \tag{3.194}$$

Bewijs. Het bewijs volgt onmiddellijk uit de inklemming van de resulterende necessiteiten tussen de grootste en kleinste marginale necessiteit die aanwezig is in P_π. □

Een eerste interessant gevolg van Stelling 3.8 is de idempotentie van $S_{\gamma^{T,F}}$:

$$\forall \widetilde{p} \in \mathcal{F}(\mathbb{B}) : S_{\gamma^{T,F}}\{\widetilde{p}, ..., \widetilde{p}\} = \widetilde{p} \tag{3.195}$$

Dit kan worden ingezien doordat voor een multiverzameling van gelijke possibilistische waarheidswaarden, zowel de Zadeh-uitbreiding van $\wedge$ als de Zadeh-uitbreiding van $\vee$, gelijk zijn aan deze waarheidswaarde. Een tweede interessant gevolg is de orthogonaliteit tussen afhankelijkheid in de zin van De Cooman en waarheidsbehoudende afhankelijkheid. Meer bepaald geldt er dat:

$$\left(\widetilde{p}_1 \,\widetilde{\wedge}_t\, ... \,\widetilde{\wedge}_t\, \widetilde{p}_n\right) \leq S_{\gamma^{T,F}}(P_\pi) \leq \left(\widetilde{p}_1 \,\widetilde{\vee}_t\, ... \,\widetilde{\vee}_t\, \widetilde{p}_n\right). \tag{3.196}$$

De volgende stelling toont aan dat het drastisch zijn van γ^F toelaat om $S_{\gamma^{T,F}}$ te herschrijven in termen van de Zadeh-uitbreiding van getransformeerde possibilistische waarheidswaarden.

Stelling 3.9 (Transformatiestelling I)
Indien γ^F drastisch is, dan geldt er:

$$S_{\gamma^{T,F}}(P_\pi) = \widetilde{\bigwedge_{i \in \{1,...,n\}}} T_S\left(\mathbf{w}_i, \widetilde{p}_{(i)^T}\right) \tag{3.197}$$

waarbij $\mathbf{w}_i = 1 - \gamma^T(A_{(i-1)^T})$ *en* T_S *een transformatiefunctie is, zodat:*

$$T_S(w, \widetilde{p}) = \begin{cases} \widetilde{p} & \textbf{\textit{als}} \quad w = 1 \\ (1, \min(w, \mu_{\widetilde{p}}(F))) & \textbf{\textit{als}} \quad w < 1. \end{cases} \tag{3.198}$$

Bewijs. (1) We tonen eerst aan dat de uitdrukkingen voor de possibiliteit van waar equivalent zijn. Het drastisch zijn van γ^F impliceert dat:

$$\begin{aligned} \exists m \in \{1, ..., n-1\} \quad & : \left(\forall k \leq m : \gamma^F(A_{(k)^F}) = 0\right) \\ & \wedge \left(\forall k > m : \gamma^F(A_{(k)^F}) = 1\right). \end{aligned} \tag{3.199}$$

Aangezien $\min\left(\gamma^T(Q), \gamma^F(\overline{Q})\right) = 0$, geldt er:

$$\forall k > m : \gamma^T(\overline{A_{(k)^F}}) = \gamma^T(A_{(n-k)^T}) = 0. \tag{3.200}$$

Dit kunnen we herschrijven als:

$$\forall k < n - m : \gamma^T(A_{(k)^T}) = 0. \tag{3.201}$$

Verder geldt er dat $\widetilde{p}_{(i)^T} = \widetilde{p}_{(n-i+1)^F}$ en vice versa. Het toepassen van de Sugeno-integraal kan dan worden geschreven als:

$$\begin{aligned} \mathrm{Nec}(p = F) &= \bigvee_{i=1}^{n} \left(\mathrm{Nec}(\mathcal{A}_{(i)^F} = F) \wedge \gamma^F(A_{(i)^F})\right) && (3.202) \\ &= \bigvee_{i=m+1}^{n} \mathrm{Nec}(\mathcal{A}_{(i)^F} = F) && (3.203) \\ &= \bigvee_{i=m+1}^{n} \mathrm{Nec}(\mathcal{A}_{(n-i+1)^T} = F) && (3.204) \\ &= \bigvee_{i=1}^{n-m} \mathrm{Nec}(\mathcal{A}_{(i)^T} = F). && (3.205) \end{aligned}$$

Herschrijven we dit in termen van possibiliteiten dan komt er:

$$\begin{aligned} \mathrm{Pos}(p = T) &= 1 - \mathrm{Nec}(p = F) && (3.206) \\ &= 1 - \bigvee_{i=1}^{n-m} \mathrm{Nec}(\mathcal{A}_{(i)^T} = F) && (3.207) \\ &= \bigwedge_{i=1}^{n-m} \mathrm{Pos}(\mathcal{A}_{(i)^T} = T). && (3.208) \end{aligned}$$

De alternatieve berekening vertrekt van een vector $\mathbf{w}$ waarbij:

$$\mathbf{w}_i = 1 - \gamma^T(A_{(i-1)^T}). \tag{3.209}$$

Gelet op:

$$\forall k \in \{1, ..., n - m - 1\} : \gamma^T(A_{(i)^T}) = 0 \tag{3.210}$$

vinden we dat:

$$\forall k \in \{1, ..., n-m\} : \mathbf{w}_i = 1. \tag{3.211}$$

Aangezien enerzijds:

$$w < 1 \Rightarrow \mu_{T_S(w,\widetilde{p})}(T) = 1 \tag{3.212}$$

en anderzijds:

$$w = 1 \Rightarrow \mu_{T_S(w,\widetilde{p})}(T) = \mu_{\widetilde{p}}(T) \tag{3.213}$$

kan de possibiliteit voor waar van de uitgebreide conjunctie van getransformeerde possibilistische waarheidswaarden worden herschreven als:

$$\begin{aligned} \text{Pos}(p = T) &= \bigwedge_{i=1}^{n} \mu_{T_S(\mathbf{w}_i, \widetilde{p}_{(i)^T})}(T) && (3.214)\\ &= \bigwedge_{i=1}^{n-m} \mu_{\widetilde{p}_{(i)^T}}(T) && (3.215)\\ &= \bigwedge_{i=1}^{n-m} \text{Pos}(\mathcal{A}_{(i)^T} = T). && (3.216) \end{aligned}$$

De uitdrukkingen voor $\text{Pos}(p = T)$ zijn bijgevolg equivalent.

(2) Vervolgens tonen we aan dat de uitdrukkingen voor de possibiliteit van vals equivalent zijn. In het geval van de Sugeno-integraal krijgen we:

$$\text{Nec}(p = T) = \bigvee_{i=1}^{n} \big(\text{Nec}(\mathcal{A}_{(i)^T} = T) \wedge \gamma^T(A_{(i)^T})\big) \tag{3.217}$$

daar waar de alternatieve berekening leidt tot:

$$\begin{aligned} \text{Pos}(p = F) &= \bigvee_{i=1}^{n} \mu_{T_S(\mathbf{w}_i, \widetilde{p}_{(i)^T})}(F) && (3.218)\\ &= \bigvee_{i=1}^{n} \big(\text{Pos}(\mathcal{A}_{(i)^T} = F) \wedge \mathbf{w}_i\big) && (3.219)\\ &= \bigvee_{i=1}^{n} \big(\big(1 - \text{Nec}(\mathcal{A}_{(i)^T} = T)\big) \wedge \big(1 - \gamma^T(A_{(i-1)^T})\big)\big) && (3.220) \end{aligned}$$

Gelet op Stelling 3.6 geldt er dat:

$$\text{Nec}(p = T) = 1 - \text{Pos}(p = F). \tag{3.221}$$

□

Een duale stelling bestaat in het geval van een drastische vertrouwensmaat γ^T.

Stelling 3.10 (Transformatiestelling II)
Indien γ^T drastisch is, dan geldt er:

$$S_{\gamma^{T,F}}(P_\pi) = \widetilde{\bigvee_{i\in\{1,\ldots,n\}}} F_S\left(\mathbf{w}_i, \widetilde{p}_{(i)^F}\right) \tag{3.222}$$

waarbij $\mathbf{w}_i = 1 - \gamma^F(A_{(i-1)^F})$ en F_S een transformatie functie is, zodat:

$$F_S(w,\widetilde{p}) = \begin{cases} \widetilde{p} & \textbf{\textit{als}} \quad w = 1 \\ \left(\min\left(w, \mu_{\widetilde{p}}(T)\right), 1\right) & \textbf{\textit{als}} \quad w < 1 \end{cases} \tag{3.223}$$

Bewijs. Het bewijs van deze tweede transformatiestelling is volledig analoog aan het bewijs van de eerste transformatiestelling. □

Stellingen 3.9 en 3.10 laten zien dat er een verband bestaat tussen $S_{\gamma^{T,F}}$ en de methode van De Tré en De Baets voor kenniscombinatie. Er bestaat een sterke analogie tussen enerzijds de transformatiefuncties T_S en F_S en anderzijds de conjunctieve transformatie g_c en de disjunctieve transformatie g_d. Meer bepaald vinden we dat:

$$\forall x \in [0,1] : \forall \widetilde{p} \in \mathcal{F}(\mathbb{B}) \quad : \quad \widetilde{p} \leq T_S\left(x, \widetilde{p}\right) \tag{3.224}$$
$$\forall x \in [0,1] : \forall \widetilde{p} \in \mathcal{F}(\mathbb{B}) \quad : \quad \widetilde{p} \geq F_S\left(x, \widetilde{p}\right) \tag{3.225}$$

hetgeen eveneens geldt voor de conjunctieve en disjunctieve transformatie van De Tré en De Baets (zie (3.35)). Als randgevallen vinden we:

$$\forall \widetilde{p} \in \mathcal{F}(\mathbb{B}) \quad : \quad T_S\left(1, \widetilde{p}\right) = \widetilde{p} \tag{3.226}$$
$$\forall \widetilde{p} \in \mathcal{F}(\mathbb{B}) \quad : \quad F_S\left(1, \widetilde{p}\right) = \widetilde{p} \tag{3.227}$$
$$\forall \widetilde{p} \in \mathcal{F}(\mathbb{B}) \quad : \quad T_S\left(0, \widetilde{p}\right) = (1,0) \tag{3.228}$$
$$\forall \widetilde{p} \in \mathcal{F}(\mathbb{B}) \quad : \quad F_S\left(0, \widetilde{p}\right) = (0,1). \tag{3.229}$$

Gelet op (3.37) vinden we equivalente randgevallen voor de conjunctieve en disjunctieve transformatie van De Tré en De Baets. Tot slot voldoen de transformatiefuncties aan de volgende eigenschap.

Eigenschap 3.2
De transformatiefuncties T_S en F_S zijn verbonden door de Zadeh-uitbreiding $\widetilde{\neg}$ van $\neg$ als volgt:

$$\forall w \in [0,1] : \forall \widetilde{p} \in \mathcal{F}(\mathbb{B}) : T_S(w,\widetilde{p}) = \widetilde{\neg}\left(F_S(w, \widetilde{\neg}\,\widetilde{p})\right). \tag{3.230}$$

Bewijs. Het bewijs volgt voor het geval $w = 1$ uit de randgevallen van de transformatiefuncties. Voor het geval $w < 1$ vinden we dat:

$$\begin{aligned} \widetilde{\neg}\, F_S(w, \widetilde{\neg}\,\widetilde{p}) &= \widetilde{\neg}\, F_S\left(w, (\mu_{\widetilde{p}}(F), \mu_{\widetilde{p}}(T))\right) && (3.231)\\ &= \widetilde{\neg}\left(\min\left(w, \mu_{\widetilde{p}}(F)\right), \mu_{\widetilde{p}}(T)\right) && (3.232)\\ &= \left(\mu_{\widetilde{p}}(T), \min\left(w, \mu_{\widetilde{p}}(F)\right)\right) && (3.233)\\ &= T_S\left(w, \widetilde{p}\right). && (3.234) \end{aligned}$$

□

Eigenschap 3.2 is het equivalent van (3.28) in de methode van De Tré en De Baets. Hoewel deze eigenschappen een duidelijk verband tonen tussen enerzijds de functie T_S en g_c en anderzijds de functie F_S en g_d, is er ook een belangrijk verschil. De beeldverzameling van de functies T_S en F_S is niet beperkt tot drie possibilistische waarheidswaarden bij constante $\widetilde{p}$, daar waar dit wel het geval is voor de functies g_c en g_d. De neutralisatie van onzekerheid kan worden gestuurd door variatie van w. Daar waar g_c en g_d kunnen afbeelden op $(1,1)$, kan deze laatste waarde nooit het beeld zijn van T_S of F_S als $\widetilde{p} \neq (1,1)$. Dit is in overeenkomst met het drastisch zijn van één van beide vertrouwensmaten, waardoor geen enkele vooropgestelde situatie kan leiden tot complete onzekerheid. Door de meer subtiele sturing van neutralisatie kunnen we besluiten dat onze aanpak de methode van De Tré en De Baets veralgemeend.

Tot slot van dit hoofdstuk bestuderen we een aantal bijzondere gevallen voor de conditionele necessiteit die interessant zullen blijken in volgende hoofdstukken. Een eerste bijzonder geval doet zich voor wanneer $\gamma^T(Q)$ en $\gamma^F(Q)$ enkel afhangen van de kardinaliteit van Q. Meer bepaald geldt er dan:

$$\forall Q_1 \subseteq A \quad : \quad \forall Q_2 \subseteq A : (|Q_1| = |Q_2|) \Rightarrow \left(\gamma^T(Q_1) = \gamma^T(Q_2)\right) \quad (3.235)$$

$$\forall Q_1 \subseteq A \quad : \quad \forall Q_2 \subseteq A : (|Q_1| = |Q_2|) \Rightarrow \left(\gamma^F(Q_1) = \gamma^F(Q_2)\right). \quad (3.236)$$

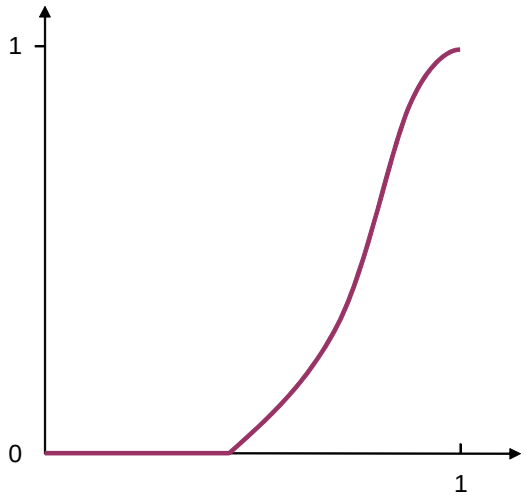
We spreken in dit geval van *kardinaliteitsgebaseerde* vertrouwensmaten. In dit geval kunnen we de vertrouwensmaten koppelen aan vage kwantoren. Vage kwantoren zijn ingevoerd door Zadeh [62] en vormen een vervaging van de twee klassieke kwantoren, namelijk de universele kwantor ($\forall$) en de existentiële kwantor ($\exists$). Zadeh stelt ondermeer dat elke relatieve[1] linguïstische uitdrukking van hoeveelheid kan worden voorgesteld door een vaagverzameling $\widetilde{V}$ over het domein $[0,1]$. Nemen we bijvoorbeeld de uitdrukking 'meeste', dan kunnen we een vaagverzameling $\widetilde{V}_{\text{meeste}}$ vooropstellen zoals afgebeeld in Figuur 3.4. Voor elke hoeveelheid $h \in [0,1]$ geeft $\widetilde{V}_{\text{meeste}}(h)$ de mate waarin h overeenstemt met het concept 'meeste'. Dergelijke kwantoren kunnen we gebruiken om conditionele necessiteit op te bouwen. Veronderstellen we in het algemeen een linguïstische uitdrukking van hoeveelheid L, die wordt gemodelleerd door een vaagverzameling $\widetilde{V}_L$ over het eenheidsinterval $[0,1]$. Beschouwen we een uitspraak van de vorm:

$$p \text{ is } b, \text{ als } L \text{ actoren stellen dat } p \text{ gelijk is aan } b.$$

met $b \in \mathbb{B}$. Bovenstaande uitspraak kan worden geëvalueerd door gebruik te maken van de functie $S_{\gamma^{T,F}}$ waarbij de vertrouwensmaten geconstrueerd worden als volgt:

$$\gamma^b(Q) = \mu_{\widetilde{V}_L}\left(\frac{|Q|}{|A|}\right). \quad (3.237)$$

[1] Relatief dient hier te worden geïnterpreteerd als verhoudingsgewijs ten opzichte van een referentiepunt.



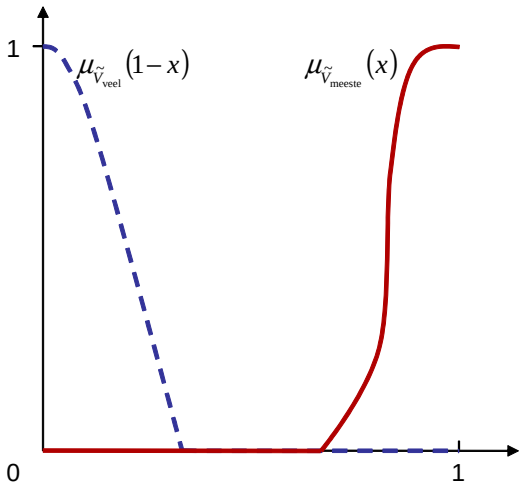
Figuur 3.4: Vaagverzameling die de vage kwantor "meeste" modelleert

Indien voor slechts één van de waarheidswaarden een linguïstische uitspraak is gedaan, veronderstellen we een drastische vertrouwensmaat voor de andere. Indien voor beide waarheidswaarden een linguïstische uitspraak is geformuleerd, kunnen beiden worden gebruikt. Echter, in dit geval moeten de vaagverzamelingen uit beide uitdrukkingen aan de volgende voorwaarde voldoen:

$$\forall x \in [0,1] : \min\left(\mu_{\widetilde{V}_{L_T}}(x), \mu_{\widetilde{V}_{L_F}}(1-x)\right) = 0 \tag{3.238}$$

waarbij L_T en L_F de linguïstische uitdrukkingen voor respectievelijk waar en vals zijn. Deze voorwaarde is nodig opdat aan (3.68) voldaan zou zijn. Wanneer twee verschillende linguïstische uitdrukkingen worden gebruikt, spreken we van bipolaire kwantificatie. Een voorbeeld van dergelijke bipolaire kwantificatie wordt getoond in Figuur 3.5.

Het begrip 'bipolair' duidt erop dat de kennis over twee tegengestelde gebeurtenissen niet noodzakelijk complementerend hoeft te zijn [63]. Vanuit dat standpunt is het gebruik van de term 'bipolair' hier inderdaad gerechtvaardigd om twee redenen. Ten eerste is voor een propositie p, de possibilistische waarheidswaarde $\widetilde{p}$ de kennis die een actor bezit over het waar en vals zijn van p. Met onze aanpak van twee vertrouwensmaten laten we toe dat het vertrouwen dat in die actor wordt gesteld, kan afhangen van de waarheidswaarde die de actor postuleert. Dit betekent dat het vertrouwen in de actor kan afhangen van het feit of hij $p = T$, dan wel $p = F$ postuleert. Ten tweede is in bipolaire systemen een bijzondere rol weggelegd voor situaties van totale onzekerheid. Het equivalent hiervan in onze aanpak is dat voor bepaalde Q, de conditionele necessiteit nul kan zijn, d.i. $\gamma^T(Q) = \gamma^F(Q) = 0$. We merken trouwens op dat dit gedrag de Sugeno-integraal voor possibilistische waarheidswaarden



Figuur 3.5: Bipolaire kwantoren

onderscheidt van de klassieke, discrete Sugeno-integraal.

We stellen nu vast dat in het geval van *kardinaliteitsgebaseerde* vertrouwensmaten een sterkere vorm van monotoniteit geldt voor $S_{\gamma^{T,F}}$. Voor een verzameling A van actoren en twee multiverzamelingen van possibilistische waarheidswaarden $P_{\pi,1}$ en $P_{\pi,2}$, zeggen we dat $P_{\pi,1}$ T-dominant is over $P_{\pi,2}$ als:

$$\forall i \in \{1, ..., n\} : \widetilde{p}_{1,(i)^T} \geq \widetilde{p}_{2,(i)^T}. \tag{3.239}$$

T-dominantie volgt steeds uit puntsgewijze dominantie maar niet omgekeerd. Wanneer de conditionele necessiteiten *kardinaliteitsgebaseerd* zijn, dan is $S_{\gamma^{T,F}}$ monotoon met betrekking tot de ordening van multiverzamelingen P_π volgens T-dominantie.

De combinatiefunctie $S_{\gamma^{T,F}}$ zal een belangrijke rol spelen in volgende hoofdstukken. In Hoofdstuk 4 zal de functie worden gebruikt in een vergelijkingsmethode voor collecties. Deze methode wordt dan gebruikt om evaluatoren voor karakterstrings te construeren (Hoofdstuk 6). In Hoofdstuk 7 zal de functie worden gebruikt om complexe objecten te vergelijken.

3.4 Conclusie

In dit hoofdstuk is de combinatie van possibilistische waarheidswaarden onderzocht. Eerst is onderzocht welke methoden hiervoor reeds bestaan. Een interessante methode in de context van coreferentiebepaling is deze van De Tré en De Baets die toelaat om onzekerheid te neutraliseren of te maximaliseren. Om deze methode te veralgemenen is een alternatieve kijk gegeven op de generatie van possibilistische waarheidswaarden. We veronderstellen in

dit hoofdstuk een Boolese propositie p waarvan we de waarheidswaarde niet kennen. We veronderstellen n actoren $\mathcal{A}_i$ die elk een uitspraak doen over p in de vorm van een possibilistische waarheidswaarde $\widetilde{p}_i$. Het probleem dat we hier hebben bestudeerd, handelt over de kennis die kan worden afgeleid over de waarheidswaarde van p, gegeven de uitspraken van de verschillende actoren. We hebben dit probleem aangepakt door het vertrouwen dat in actoren kan worden gesteld in rekening te brengen. Dit vertrouwen modelleren we aan de hand van twee vertrouwensmaten die we samen de conditionele necessiteit noemen. Vervolgens hebben we aangetoond hoe op basis van conditionele necessiteit een possibilistische waarheidswaarde kan worden afgeleid die de onzekerheid over de waarheidswaarde van de propositie p voorstelt. De combinatiefunctie die deze possibilistische waarheidswaarde genereert, vertoont een opvallende gelijkenis met de discrete Sugeno-integraal. Tot slot worden de eigenschappen van de nieuwe combinatiefunctie onderzocht. Het is aangetoond hoe de nieuwe combinatiefunctie zich verhoudt tot bestaande methoden uit de literatuur. De nieuwe combinatiefunctie zal veelvuldig worden gebruikt in de uitwerking van syntactische evaluatoren in het vervolg van deze thesis.

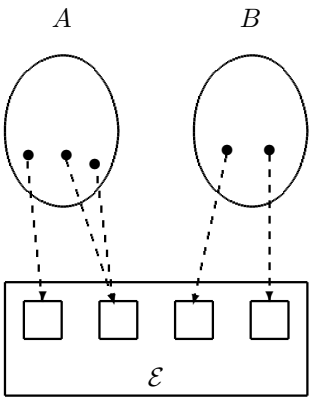
Hoofdstuk 4

Evaluatoren voor collecties

4.1 Inleiding

In dit hoofdstuk behandelen we syntactische evaluatoren in de context van een atomair universum, waarbij elementen uit dit universum collecties zijn. Met een collectie wordt hier een verzameling of een multiverzameling bedoeld. Het vergelijken van collecties heeft in de context van coreferentie verschillende toepassingen. Ten eerste kunnen collecties worden gebruikt om eigenschappen te beschrijven. Denken we bijvoorbeeld aan de kinderen van een persoon of de hobby's van een persoon. In een recente context van sociale netwerksites, waarbij profielen bestaan uit een aantal persoonlijke gegevens en een verzameling van contactpersonen, kan het toetsen van collecties op coreferentie bijzonder nuttig zijn. Wanneer men bijvoorbeeld in kaart wil brengen in welke mate personen op verschillende sociale netwerksites zijn vertegenwoordigd, moet men zoeken naar coreferente profielen. Een tweede toepassing is het vergelijken van ontologieën. Een ontologie is een netwerkstructuur die bestaat uit een collectie van concepten en een aantal binaire relaties over de verzameling van concepten. Het doel van een ontologie is het modelleren van kennis over een welbepaald domein. Indien men wil weten of twee ontologieën eenzelfde domein behandelen, dient men deze te vergelijken en te controleren op coreferente concepten en relaties. Een derde toepassing is het geval waarbij een object met behulp van een transformatiefunctie kan worden omgevormd tot een collectie van objecten. Dit principe zal in het volgende hoofdstuk worden gebruikt om evaluatoren voor karakterstrings te bekomen door karakterstrings om te zetten naar een collectie van deelstrings.

In Sectie 4.2 wordt het vergelijken van verzamelingen eerst nader toegelicht en wordt een kort overzicht gegeven van de literatuur omtrent dit probleem. Daarna wordt een evaluator voor verzamelingen gedefinieerd en wordt stapsgewijs toegelicht hoe een dergelijke evaluator werkt. Hierbij wordt gedeeltelijk gesteund op de resultaten uit Hoofdstuk 3. In Sectie 4.3 wordt aangetoond hoe de evaluator voor verzamelingen kan worden uitgebreid naar een evalua-



Figuur 4.1: Twee verzamelingen van objecten en de entiteitsbeschrijving van hun elementen.

tor voor multiverzamelingen. Om de praktische toepasbaarheid te garanderen, wordt een complexiteitsanalyse gemaakt in Sectie 4.4. Ten slotte worden enkele eigenschappen van de voorgestelde aanpak onderzocht in Sectie 4.5 met het oog op de verdere toepassing ervan in Hoofdstuk 6, dat handelt over evaluatoren voor karakterstrings. In Sectie 4.6 wordt het verband met gedeeltelijke coreferentie getoond. Sectie 4.7 vat de belangrijkste resultaten van dit hoofdstuk samen.

4.2 Coreferentie van verzamelingen

Het vergelijken van verzamelingen wordt in de literatuur vaak opgelost door elementen te vergelijken op basis van de gelijkheidsrelatie $=$. Dit leidt tot oplossingen die een evaluatie maken van afgeleide verzamelingen zoals de doorsnede, de unie en het verschil. De oudste en wellicht meest gekende methode volgens dit principe is deze van Jaccard [64], waarbij de mate van overeenkomst tussen twee verzamelingen gelijk is aan de verhouding van de kardinaliteit van de doorsnede tot de kardinaliteit van de unie.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}. \tag{4.1}$$

Het werk van Jaccard vormt de basis van een uitgebreid onderzoeksgebied [65, 66] dat ook in termen van vaagverzamelingen diepgaand is onderzocht [67, 68, 69]. In onze aanpak voor het vergelijken van collecties willen we expliciet in rekening brengen dat gelijkheid en coreferentie van objecten niet altijd equivalent zijn. Daarom veronderstellen we dat verzamelingen coreferent zijn, als ze bestaan uit coreferente elementen, eerder dan gelijke elementen. Hierdoor wordt er rekening gehouden met onzekerheid over coreferentie op het niveau van elementen. Een overzicht van oorzaken van onzekerheid bij coreferentiebepaling is gegeven in Hoofdstuk 2. In het kader van dit overzicht stellen

we vast dat sommige van deze oorzaken[1] kunnen leiden tot verschillen in de kardinaliteiten van coreferente verzamelingen. Een verschil in kardinaliteit zal bijgevolg niet altijd wijzen op het niet-coreferent zijn van collecties. Als we bijvoorbeeld aan twee verschillende personen vragen wie de vrienden zijn van een derde persoon, zullen de resulterende verzamelingen niet noodzakelijk gelijk zijn in kardinaliteit. In het kader van dit probleem wordt opgemerkt dat ook de kardinaliteit van gerefereerde entiteiten in rekening moet worden gebracht. Stel bijvoorbeeld een verzameling van objecten A waarbij sommige van deze objecten onderling coreferent zijn. In dit geval is het aantal verschillende entiteiten beschreven door de objecten in A kleiner dan $|A|$. Wanneer we A met een andere verzameling B vergelijken, willen we in kaart brengen dat A exact $|A|$ entiteiten beschrijft, zij het één of meer entiteiten meer dan één keer (zie Figuur 4.1). De multipliciteit waarmee entiteiten worden beschreven door objecten in collecties is dan van belang bij het beantwoorden van de vraag: "Zijn twee collecties coreferent?". Het model voor vergelijking van collecties kan dan worden omschreven als volgt:

Twee verzamelingen zijn coreferent als ze
dezelfde entiteiten beschrijven.

In het vervolg veronderstellen we twee verzamelingen van objecten A en B die elementen bevatten uit een universum U, waarbij er zonder de algemeenheid te schaden, geldt dat $|A| \leq |B|$. De propositie die stelt dat A en B coreferente verzamelingen zijn, wordt geëvalueerd door een evaluator $E_{\mathcal{P}(U)}$. Deze evaluatie gebeurt in twee stappen. In een eerste stap, wordt een evaluator E_U over het universum U verondersteld. Deze evaluator wordt gebruikt voor de constructie van een injectieve afbeelding ι van A naar B. In een tweede stap worden de koppels van elementen in ι geëvalueerd door E_U, waardoor een sequentie van possibilistische waarheidswaarden wordt geproduceerd. De propositie "A en B zijn coreferent" kan dan worden geëvalueerd door combinatie van deze possibilistische waarheidswaarden. Als combinatiefunctie gebruiken we de Sugeno-integraal voor possibilistische waarheidswaarden uit Hoofdstuk 3.

Definitie 4.1 (Evaluator voor verzamelingen)
Veronderstel een universum U. Een evaluator voor verzamelingen van elementen uit U, gebaseerd op een evaluator E_U, is gedefinieerd als:

$$E_{\mathcal{P}(U)} : \mathcal{P}(U)^2 \to \mathcal{F}(\mathbb{B}) : (A, B) \mapsto S_{\gamma^{T,F}}(P^{\iota}_{\pi,(A,B)}) \tag{4.2}$$

met $P^{\iota}_{\pi,(A,B)}$ een multiverzameling van $|B|$ possibilistische waarheidswaarden die moet voldoen aan:

$$\forall (a, b) \in \iota : E_U(a, b) \in P^{\iota}_{\pi,(A,B)} \tag{4.3}$$

en waarbij bijkomend $|B| - |A|$ possibilistische waarheidswaarden worden toegevoegd die gelijk zijn aan $(0, 1)$.

[1] In het bijzonder ruis, onnauwkeurigheid en het niet-eenduidig bepaald zijn van het meetresultaat.

In wat volgt bespreken we eerst de constructie van ι. Daarna zal op basis van de eigenschappen van ι, het combineren van de possibilistische waarheidswaarden worden besproken.

4.2.1 Generatie van een injectieve afbeelding

Het gebruik van een injectieve afbeelding tussen twee verzamelingen wordt verklaard door het ingevoerde model voor vergelijking van verzamelingen. Dit model stelt dat coreferente verzamelingen dezelfde entiteiten beschrijven. Om dit te bewerkstelligen wordt elke gerefereerde entiteit als een afzonderlijke eenheid beschouwd. Als een entiteit bijvoorbeeld door twee objecten in verzameling A wordt beschreven en door drie objecten in verzameling B, dan kunnen we dit verschil in kaart brengen door elk van de twee objecten uit A af te beelden op precies één van de drie objecten uit B, waarbij de beelden van beide elementen uit A verschillend zijn. Het vooropgestelde model wordt bijgevolg bereikt als elementen één aan één worden gekoppeld. Dit kan door een injectieve afbeelding van A (kleinste kardinaliteit) naar B (grootste kardinaliteit) te construeren. Merk op dat hieruit volgt dat ι precies $|A|$ koppels moet bevatten.

In wat volgt zullen we voor de combinatiefunctie $S_{\gamma^{T,F}}$ veronderstellen dat γ^F drastisch is[2]. Dit betekent dat de beeldverzameling van γ^F wordt beperkt tot twee waarden én dat γ^F volledig afhankelijk is van γ^T. De nadruk ligt bijgevolg op de bepaling van γ^T, hetgeen wil zeggen dat we de zekerheid over coreferentie van collecties laten afhangen van de zekerheid over coreferentie van de elementen. Dit betekent dat we voor de constructie van de afbeelding ι iteratief op zoek gaan naar het koppel van elementen waarvoor de zekerheid op coreferentie het hoogst is. Vermits er expliciet wordt verondersteld dat E_U sterk reflexief is[3], volgt voor elementen in $A \cap B$ onmiddellijk dat zij onder ι op zichzelf afgebeeld moeten worden. Voor de verdere constructie van ι beschouwen we de matrix $\mathbf{M}_{A,B}$ van possibilistische waarheidswaarden waarbij de rijen geïndexeerd zijn door elementen uit $A \backslash B$ en de kolommen door elementen uit $B \backslash A$. Uit deze matrix zullen we iteratief de grootste possibilistische waarheidswaarde extraheren en vervolgens de elementen gekoppeld aan de rij en kolom uit $\mathbf{M}_{A,B}$ waar deze possibilistische waarheidswaarde staat, toevoegen aan ι. De grootste overblijvende possibilistische waarheidswaarde noteren we als $\widetilde{p}_{\max}$ en wordt gegeven door:

$$\widetilde{p}_{\max} = \widetilde{\bigvee_{(a,b) \in A_\iota \times B_\iota}} \mathbf{M}_{A,B}(a,b) \tag{4.4}$$

waarbij de volgende notaties worden gebruikt:

$$A_\iota = \{a | a \in A \wedge (\forall b \in B : (a,b) \notin \iota)\} \tag{4.5}$$

$$B_\iota = \{b | b \in B \wedge (\forall a \in A : (a,b) \notin \iota)\}. \tag{4.6}$$

[2]Het zal later blijken dat deze veronderstelling logisch voortkomt uit een redenering met beslissingsmodellen. Hiervoor verwijzen we naar Hoofdstukken 5 en 7.

[3]Dit gegeven verandert fundamenteel niets aan onze aanpak, maar is belangrijk om redenen van complexiteit.

Wanneer $\widetilde{p}_{\max}$ op meerdere posities in de matrix $\mathbf{M}_{A,B}$ voorkomt, moet een keuze worden gemaakt. Tabel 4.1 toont een voorbeeld van het keuzeprobleem dat zich kan stellen. In dit geval is $\widetilde{p}_{\max}$ gelijk aan $(1, 0.1)$ en deze possibilistische waarheidswaarde komt twee maal voor. Dit stelt ons voor de keuze a af te beelden op c of b af te beelden op c. Afbeelding van b op c impliceert dat een grotere possibilistische waarheidswaarde overblijft, aangezien $(1, 0.3) > (1, 0.7)$. Onze voorkeur gaat dus uit naar afbeelding van b op c.

	c	d
a	(1,0.1)	(1,0.3)
b	(1,0.1)	(1,0.7)

Tabel 4.1: Keuzeprobleem bij constructie van ι

De voorkeur die we in dit voorbeeld uitdrukken kan worden geformaliseerd door te zeggen dat de grootste possibilistische waarheidswaarde uit de matrix moet worden gehaald, met als randvoorwaarde dat de volledige sequentie van possibilistische waarheidswaarden maximaal is onder de leximax orderelatie. Deze orderelatie wordt voor een vector van possibilistische waarheidswaarden gedefinieerd als volgt [70].

Definitie 4.2 (Leximax orderelatie)
Voor twee vectoren van possibilistische waarheidswaarden $\mathbf{x} \in \mathcal{F}(\mathbb{B})^n$ *met* $\mathbf{x}_i \geq \mathbf{x}_{i+1}$ *en* $\mathbf{y} \in \mathcal{F}(\mathbb{B})^n$ *met* $\mathbf{y}_i \geq \mathbf{y}_{i+1}$ *is de leximax orderelatie gedefinieerd als:*

$$\mathbf{x} <_{\text{leximax}} \mathbf{y} \Leftrightarrow \exists k \in \{1, ..., n\} : \mathbf{x}_k < \mathbf{y}_k \wedge (\forall l < k : \mathbf{x}_l = \mathbf{y}_l) \tag{4.7}$$

Stel nu dat $\widetilde{p}_{\max}$ op k plaatsen in de matrix $\mathbf{M}_{A,B}$ voorkomt. Dit wil zeggen dat er k koppels bestaan waarvoor de evaluatie onder E_U gelijk is aan $\widetilde{p}_{\max}$. Laat ons deze koppels noteren als $\{(a_1, b_1), ..., (a_k, b_k)\}$. Voor elk koppel (a_i, b_i) uit deze verzameling kunnen we verifiëren wat $\widetilde{p}_{\max}$ wordt ná toevoeging van (a_i, b_i) aan ι. Vermits ι moet leiden tot een sequentie van possibilistische waarheidswaarden die zo groot mogelijk is onder de leximax orderelatie, moet $\widetilde{p}_{\max}$ na toevoeging van (a_i, b_i) aan ι zo groot mogelijk zijn. Koppels uit $\{(a_1, b_1), ..., (a_k, b_k)\}$ waarvoor de toevoeging aan ι tot gevolg heeft dat de nieuwe $\widetilde{p}_{\max}$ kleiner is dan voor een willekeurig ander koppel uit $\{(a_1, b_1), ..., (a_k, b_k)\}$, mogen worden verwijderd uit deze verzameling. In termen van een algoritme kunnen we zeggen dat de lijst van mogelijke koppels kan worden onderzocht met een breedte-eerste zoekstrategie. Deze strategie wordt in pseudocode gegeven door Algoritme 4.1.

Algoritme 4.1 begint met de constructie van ι door elementen in $A \cap B$ op zichzelf af te beelden onder ι (regel 1). Vervolgens worden er koppels van elementen aan ι toegevoegd zolang ι minder koppels bevat dan er elementen in A zijn. Deze toevoeging gebeurt door eerst de grootste possibilistische waarheidswaarde te zoeken (regel 3) en daarna de beschikbare koppels te selecteren waarvoor de evaluatie onder E_U gelijk is aan $\widetilde{p}_{\max}$ (regel 4). Voor elk van deze koppels wordt een afbeelding ι_i voorzien (regel 5). De procedure **choose** kiest

Algoritme 4.1 mapping(A, B)

1: $\forall u \in (A \cap B) : \iota \leftarrow \iota \cup \{(u, u)\}$
2: **while** $|\iota| \neq |A|$ **do**
3: $\quad \widetilde{p}_{\max} \leftarrow \widetilde{\bigvee}_{(a,b)\in A_\iota \times B_\iota} \mathbf{M}_{A,B}(a, b)$
4: $\quad K \leftarrow \{(a, b) | (a, b) \in (A_\iota \times B_\iota) \wedge E_U(a, b) = \widetilde{p}_{\max}\}$
5: $\quad \forall (a_i, b_i) \in K : \iota_i = \iota \cup \{(a_i, b_i)\}$
6: $\quad \iota \leftarrow$ **choose**$(\{\iota_1, ..., \iota_{|K|}\})$
7: **end while**
8: **return** ι

uit de mogelijke afbeeldingen deze die optimaal is onder de leximax orderelatie. De strategie achter dit keuzemechanisme wordt in pseudocode gegeven door Algoritme 4.2.

Algoritme 4.2 choose$(\{\iota_1, ..., \iota_k\})$

1: $K^* \leftarrow \emptyset$
2: $\widetilde{p}^*_{\max} \leftarrow \widetilde{\bigvee}_{j\in\{1,...,k\}} \left(\widetilde{\bigvee}_{(a,b)\in(A_{\iota_j} \times B_{\iota_j})} \mathbf{M}_{A,B}(a, b) \right)$
3: **for** $i = 1$ **to** k **do**
4: $\quad \widetilde{p} \leftarrow \widetilde{\bigvee}_{(a,b)\in(A_{\iota_i} \times B_{\iota_i})} \mathbf{M}_{A,B}(a, b)$
5: $\quad$ **if** $\widetilde{p} = \widetilde{p}^*_{\max}$ **then**
6: $\quad\quad \forall (a, b) \in (A_{\iota_i} \times B_{\iota_i}) \wedge (E_U(a, b) = \widetilde{p}) : K^* \leftarrow K^* \cup \{\iota_i \cup \{(a, b)\}\}$
7: $\quad$ **end if**
8: **end for**
9: **if** $(K^* = \{\iota^*_1\} \vee ... \vee K^* = \{\iota^*_k\}) \vee (\forall \iota_j \in K^* : |\iota_j| = |A|)$ **then**
10: $\quad$ **return** ι_1
11: **else**
12: $\quad$ **return choose**(K^*)
13: **end if**

Het kiezen tussen afbeeldingen start met het zoeken naar de grootste possibilistische waarheidswaarde die overblijft na het kiezen van één van de afbeeldingen. Deze waarheidswaarde wordt $\widetilde{p}^*_{\max}$ genoemd (regel 2). Vervolgens wordt voor elke afbeelding ι_i, waarvoor de maximaal overgebleven possibilistische waarheidswaarde gelijk is aan $\widetilde{p}^*_{\max}$, een reeks van nieuwe afbeeldingen geconstrueerd die elk één koppel meer bevatten dan ι_i. De reden dat meerdere afbeeldingen vanuit ι_i geconstrueerd kunnen worden, is dat $\widetilde{p}^*_{\max}$ opnieuw op meerdere posities in de matrix kan voorkomen. Op deze manier krijgen we een nieuwe verzameling van afbeeldingen die in termen van kardinaliteit groter zijn. Wanneer deze nieuwe verzameling een singleton is, wordt de betreffende afbeelding weergegeven (regel 9). Wanneer alle mogelijke afbeeldingen compleet zijn (d.i. evenveel koppels bevatten als er elementen zijn in A), dan geven alle overgebleven afbeeldingen aanleiding tot een leximax-optimale sequentie van possibilistische waarheidswaarden (regel 9). In elk ander geval is het keuzeprobleem nog niet opgelost en wordt er verder gezocht.

Voorbeeld 4.1
Tabel 4.2 geeft de mogelijkheden voor coreferentie tussen de elementen van twee verzamelingen $A = \{a_1, a_2, a_3\}$ *en* $B = \{b_1, b_2, b_3, b_4, b_5\}$. *Voor de constructie*

	b_1	b_2	b_3	b_4	b_5
a_1	(1,0.1)	(1,0.8)	(1,0.6)	(1,0.1)	(1,0.8)
a_2	(1,0.8)	(1,0.9)	(1,0.2)	(1,0.7)	(1,0.4)
a_3	(1,0.1)	(1,0.5)	(1,0.1)	(1,0.1)	(1,0.4)

Tabel 4.2: Voorbeeld van $\mathbf{M}_{A,B}$

van de afbeelding ι *tussen deze twee verzamelingen bepalen we eerst* $\widetilde{p}_{\max}$. *In dit geval is* $\widetilde{p}_{\max} = (1, 0.1)$ *en deze waarde komt op vijf verschillende plaatsen voor in de matrix. Dit geeft aanleiding tot vijf verschillende afbeeldingen:*

$$\iota_1 = \{(a_1, b_1)\} \tag{4.8}$$
$$\iota_2 = \{(a_1, b_4)\} \tag{4.9}$$
$$\iota_3 = \{(a_3, b_1)\} \tag{4.10}$$
$$\iota_4 = \{(a_3, b_3)\} \tag{4.11}$$
$$\iota_5 = \{(a_3, b_4)\}. \tag{4.12}$$

Uit deze afbeeldingen moet één afbeelding worden gekozen en deze keuze wordt gemaakt door Algoritme 4.2. Hiervoor wordt eerst de grootst mogelijke possibilistische waarheidswaarde berekend die overblijft na keuze van één van de afbeeldingen. We vinden in dit geval:

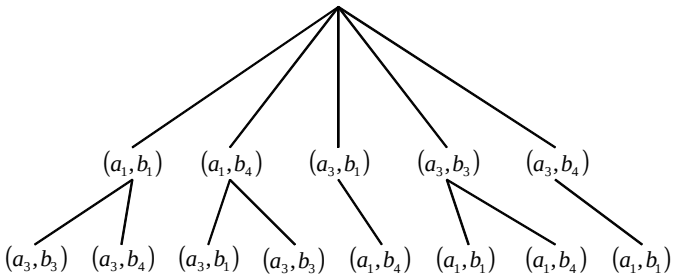
$$\widetilde{p}^*_{\max} = (1, 0.1). \tag{4.13}$$

Merk op dat voor elk van de vijf afbeeldingen, de grootst overblijvende possibilistische waarheidswaarde gelijk is aan $(1, 0.1)$. *Onder enkele van deze afbeeldingen komt* $\widetilde{p}^*_{\max}$ *op verschillende locaties voor. Dit is bijvoorbeeld zo voor* ι_1*: na afbeelding van* a_1 *op* b_1*, kan* a_3 *zowel op* b_3 *als op* b_4 *worden afgebeeld. Alle mogelijke paden in de zoekruimte worden getoond in Figuur 4.2. Het eerste niveau in deze boom correspondeert met de vijf afbeeldingen waaruit we moeten kiezen. Het tweede niveau toont de koppels die door elk van deze afbeeldingen worden toegevoegd.*

De nieuwe verzameling van afbeeldingen ziet er uit als volgt:

$$\{(a_1, b_1), (a_3, b_3)\} \tag{4.14}$$
$$\{(a_1, b_1), (a_3, b_4)\} \tag{4.15}$$
$$\{(a_1, b_4), (a_3, b_1)\} \tag{4.16}$$
$$\{(a_1, b_4), (a_3, b_3)\}. \tag{4.17}$$

Merk op dat de zoekboom acht paden bevat, maar dat deze slechts overeenstemmen met vier verschillende afbeeldingen. Aangezien nog meer dan één afbeelding in aanmerking komt en gelet op het feit dat deze afbeeldingen nog niet



Figuur 4.2: Zoekboom voor constructie van een leximax-optimale afbeelding ι

compleet zijn, worden deze vier afbeeldingen verder onderzocht door een iteratieve uitvoering van de zoekprocedure. Van de vier afbeeldingen die overblijven, zijn er slechts twee die aanleiding geven tot een leximax-optimale afbeelding, namelijk $\{(a_1, b_1), (a_3, b_4)\}$ *en* $\{(a_1, b_4), (a_3, b_1)\}$*. Voor deze afbeeldingen kan* a_2 *op* b_3 *worden afgebeeld en* $E_U(a_2, b_3)$ *is gelijk aan* $(1, 0.2)$*. Voor de andere twee afbeeldingen zou* a_2 *op* b_5 *moeten worden afgebeeld en aangezien* $E_U(a_2, b_5) = (1, 0.4) < (1, 0.2)$*, zijn deze afbeeldingen niet leximax-optimaal. Er blijven dan twee mogelijke afbeeldingen over:*

$$\{(a_1, b_1), (a_3, b_4), (a_2, b_3)\} \tag{4.18}$$

$$\{(a_1, b_4), (a_3, b_1), (a_2, b_3)\}. \tag{4.19}$$

Beide afbeeldingen zijn compleet en dus evenwaardig. Er wordt dan één van deze afbeeldingen gegeven als resultaat.

De afbeelding ι zoals geconstrueerd door Algoritme 4.1 heeft enkele interessante eigenschappen.

Eigenschap 4.1 (Symmetrie van de constructie)
De constructie van ι *is symmetrisch. Dit betekent dat er geldt:*

$$\textbf{\textit{mapping}}(A, B) = \textbf{\textit{mapping}}(B, A) \tag{4.20}$$

Bewijs. Het bewijs volgt triviaal uit het feit dat selectie van het maximum uit de matrix onafhankelijk is van de volgorde van de argumenten van **mapping**. □

Definitie 4.3 (Pareto-efficiënte afbeelding)
Voor een universum U *met evaluator* E_U *is een één-op-één afbeelding* f *tussen twee verzamelingen* $A \subset U$ *en* $B \subset U$ *Pareto-efficiënt als voor elke andere één-op-één afbeelding* g *tussen* A *en* B *geldt:*

$$\exists (a, b) \in f : E_U(a, b) \geq E_U(a, g(a)). \tag{4.21}$$

Eigenschap 4.2 (Pareto efficiëntie van ι)
De afbeelding ι *is een Pareto-efficiënte afbeelding.*

Bewijs. De Pareto-efficiëntie van ι is een onmiddellijk gevolg van het feit dat de possibilistische waarheidswaarden horend bij ι leximax-optimaal zijn. □

In het kader van bestaande methoden rond similariteit van verzamelingen [71, 72] op basis van similariteit van elementen wordt eveneens een afbeelding gebruikt, typisch met een globaal optimalisatiecriterium zoals maximalisatie van de som van de individuele similariteiten [73]. Hoewel het sommeren van possibiliteiten geen steek houdt, kan worden ingezien dat ι zoals hier beschreven niet voortkomt uit een globale optimalisatie. Immers wordt de afbeelding zo geconstrueerd dat de lokale mogelijkheid voor waar (d.i. per koppel) wordt gemaximaliseerd. Een possibilistisch equivalent van een globale aanpak zou een leximin-optimale afbeelding zijn, die ervoor zorgt dat de kleinste possibilistische waarheidswaarde zo groot mogelijk is. Een dergelijke afbeelding zou dan meteen $\widetilde{\wedge}$ -efficiënt zijn:

Definitie 4.4 ($\widetilde{\wedge}$ -efficiëntie)
Voor een universum U met evaluator E_U is een één-op-één afbeelding f tussen twee verzamelingen $A \subset U$ en $B \subset U$ $\widetilde{\wedge}$ -efficiënt als voor elke andere één-op-één afbeelding g tussen A en B geldt:

$$\widetilde{\bigwedge_{(a,b)\in f}} (E_U(a,b)) \geq \widetilde{\bigwedge_{(a,b)\in g}} (E_U(a,b)) \tag{4.22}$$

Merk op dat een leximax-optimale afbeelding niet noodzakelijk $\widetilde{\wedge}$ -efficiënt is. Dit wordt geïllustreerd aan de hand van Tabel 4.3. De leximax-optimale afbeelding $\{(a,c),(b,d)\}$ is hierbij duidelijk niet $\widetilde{\wedge}$ -efficiënt. De afbeelding $\{(a,d),(b,c)\}$ is dit wel. De voorwaarden waaronder leximax-efficiëntie leidt tot $\widetilde{\wedge}$ -efficiëntie zijn overigens streng. Transitiviteit van E_U is bijvoorbeeld geen garantie dat leximax-efficiëntie leidt tot $\widetilde{\wedge}$ -efficiëntie.

	c	d
a	(1,0.1)	(1,0.2)
b	(1,0.2)	(1,0.7)

Tabel 4.3: Leximax-efficiëntie versus $\widetilde{\wedge}$ -efficiëntie

Lokale optimalisatie in de context van similariteit wordt eveneens voorgesteld in [74, 75], maar in tegenstelling tot onze aanpak houden deze methoden geen rekening met keuzeproblemen die kunnen optreden tijdens de constructie van de afbeelding, waardoor de resulterende similariteiten niet noodzakelijk leximax-optimaal zijn en waardoor het eindresultaat ook niet noodzakelijk uniek is.

4.2.2 Kenniscombinatie

De afbeelding $\iota \subset A \times B$, zoals hierboven geconstrueerd, koppelt elementen aan elkaar, waarvoor de zekerheid op coreferentie zo groot mogelijk is. Dergelijke

koppels van elementen bieden een zekere evidentie voor het coreferent zijn van de verzamelingen waartoe de elementen behoren. Voor twee verzamelingen A en B kunnen we nu een multiverzameling $P^{\iota}_{\pi,(A,B)}$ van $|B|$ possibilistische waarheidswaarden construeren die voldoet aan:

$$\forall (a,b) \in \iota : E_U(a,b) \in P^{\iota}_{\pi,(A,B)} \tag{4.23}$$

en waarbij bijkomend $|B| - |A|$ possibilistische waarheidswaarden worden toegevoegd die gelijk zijn aan $(0,1)$. Het kan worden aangetoond dat $P^{\iota}_{\pi,(A,B)}$ uniek is voor twee willekeurige verzamelingen A en B.

Stelling 4.1
*Voor twee verzamelingen $A \subset U$ en $B \subset U$, geeft ι=**mapping**(A,B) aanleiding tot een unieke multiverzameling $P^{\iota}_{\pi,(A,B)}$ van possibilistische waarheidswaarden.*

Bewijs. Het waar zijn van deze stelling volgt uit het feit dat Algoritme 4.1 een leximax-optimale afbeelding genereert. Als er meerdere afbeeldingen bestaan, geven deze allemaal aanleiding tot dezelfde collectie van possibilistische waarheidswaarden. □

We kunnen deze multiverzameling van possibilistische waarheidswaarden nu beschouwen als de kennis die door $|B|$ actoren wordt gegenereerd. Elke actor geeft kennis over de propositie p die stelt dat verzamelingen A en B coreferent zijn. Aangezien een verschil in kardinaliteit tussen twee verzamelingen niet noodzakelijk leidt tot niet-coreferentie van die verzamelingen, betekent dit dat niet alle actoren $p = T$ moeten postuleren om te besluiten dat p waar is. Gelet op de resultaten uit Hoofdstuk 3 stellen we vast dat de nieuwe methode voor combinatie van possibilistische waarheidswaarden hiervoor geschikt is. Een belangrijke vaststelling is dat enkel de resulterende multiverzameling van possibilistische waarheidswaarden uniek is en niet de afbeelding ι waarop de multiverzameling is gebaseerd. Dit wordt geïllustreerd in Voorbeeld 4.1. Een gevolg hiervan is dat de conditionele necessiteit niet mag afhangen van de actoren die kennis genereren. Wel mag de conditionele necessiteit afhangen van het aantal actoren dat een bepaald postulaat formuleert. In Hoofdstuk 3 is een bijzonder geval van de functie $S_{\gamma^T,F}$ besproken waarbij γ^F drastisch is en waarbij γ^T is gebaseerd op een vage kwantor. De vertrouwensmaten γ^T en γ^F worden dan *kardinaliteitsgebaseerd* genoemd. Een dergelijke functie kan hier worden gebruikt, aangezien de conditionele necessiteiten γ^T en γ^F dan onafhankelijk zijn van de koppels elementen in ι en bijgevolg ook van de actoren die kennis genereren. Stelling 3.9 toont aan dat het resultaat van $S_{\gamma^T,F}$ onder deze voorwaarden kan worden geschreven in termen van de Zadeh-uitbreiding van $\wedge$, toegepast op getransformeerde possibilistische waarheidswaarden. Voor twee verzamelingen A en B (met $|A| \leq |B|$) wordt de onzekerheid omtrent coreferentie dan gegeven door de possibilistische waarheidswaarde:

$$\widetilde{\bigwedge_{i \in \{1,\ldots,|B|\}}} T_S\left(\mathbf{w}_i, \widetilde{p}_{(i)^T}\right) \tag{4.24}$$

waarbij **w** een gewichtsvector is die de conditionele necessiteit voorstelt. Het gebruik van kwantoren wordt verder bestudeerd in Sectie 4.5.

4.3 Coreferentie van multiverzamelingen

Tot hier toe hebben we een evaluator voor verzamelingen geïntroduceerd. Deze evaluator kan nu worden veralgemeend voor het geval van multiverzamelingen. In het model voor vergelijking van verzamelingen is impliciet rekening gehouden met het feit dat een verzameling coreferente elementen (d.z. objecten) kan bevatten. Dit geeft aanleiding tot de één-op-één afbeelding ι, die er voor zorgt dat voor elke entiteit het aantal refererende objecten in kaart wordt gebracht. De overgang van verzamelingen naar multiverzamelingen heeft tot gevolg dat een element meerdere keren kan voorkomen. Twee gelijke elementen zijn echter twee elementen die met zekerheid coreferent zijn. Om die reden kan het geval van multiverzamelingen volledig analoog worden behandeld en behoeven de algoritmen beschreven in de vorige sectie geen aanpassing.

Definitie 4.5 (Evaluator voor multiverzamelingen)
Veronderstel een universum U. Een evaluator voor multiverzamelingen van elementen in U, gebaseerd op een evaluator E_U, is gedefinieerd als:

$$E_{\mathcal{M}(U)} : \mathcal{M}(U)^2 \to \mathcal{F}(\mathbb{B}) : (A, B) \mapsto S_{\gamma^{T,F}}(P^{\iota}_{\pi,(A,B)}) \tag{4.25}$$

met $P^{\iota}_{\pi,(A,B)}$ een multiverzameling van $|B|$ possibilistische waarheidswaarden die moet voldoen aan:

$$\forall (a, b) \in \iota : E_U(a, b) \in P^{\iota}_{\pi,(A,B)} \tag{4.26}$$

en waarbij bijkomend $|B| - |A|$ possibilistische waarheidswaarden worden toegevoegd die gelijk zijn aan $(0, 1)$.

De uitbreiding naar het geval van multiverzamelingen zal worden gebruikt in het volgende hoofdstuk voor de constructie van evaluatoren voor karakterstrings.

4.4 Complexiteitsanalyse

In dit deel wordt de complexiteit van het vergelijken van (multi)verzamelingen besproken en worden enkele suggesties gedaan om de complexiteit te reduceren. Voor de constructie van ι moet eerst een matrix van possibilistische waarheidswaarden worden berekend. Deze stap heeft een kwadratische complexiteit in termen van de kardinaliteiten van de verschilverzamelingen. Meer bepaald heeft de berekening van de possibilistische waarheidswaarden een complexiteit:

$$O(|A \backslash B||B \backslash A|C(E_U)) \tag{4.27}$$

waarbij $C(E_U)$ de complexiteit van de evaluator E_U voorstelt. Deze afhankelijkheid in complexiteit vormt een belangrijk uitgangspunt in Hoofdstuk 6. De berekening van de matrix kan worden versneld als E_U een transitieve evaluator is (Hoofdstuk 2). Indien voor een verzameling B de mogelijkheden van coreferentie voor koppels van objecten uit B gekend zijn, dan kan de transitiviteit van E_U worden gebruikt om overtollige berekeningen te vermijden. Zo leidt de beperking op de necessiteit van waar tot de regel:

$$\begin{array}{lcl} \forall(a,b,c) \in U^3 & : & E_U(a,b) > (1,1) \\ & \wedge & E_U(b,c) > (1,1) \\ & \wedge & E_U(a,b) \neq E_U(b,c) \\ & \Rightarrow & E_U(a,c) = E_U(a,b) \mathbin{\widetilde{\wedge}} E_U(b,c). \end{array} \tag{4.28}$$

In gevallen van volledige zekerheid geldt de regel:

$$\begin{array}{lcl} \forall(a,b,c) \in U^3 & : & E_U(a,b) = (1,0) \\ & \wedge & E_U(b,c) = (0,1) \\ & \Rightarrow & E_U(a,c) = (0,1). \end{array} \tag{4.29}$$

Indien een transitieve evaluator wordt gebruikt, kan de reductie in complexiteit voor $E_{\mathcal{M}(U)}$ aanzienlijk zijn als de complexiteit van E_U hoog is. Dat het voorzien van een transitieve evaluator meestal geen evidente zaak is, hoeft niet te betekenen dat dit principe niet toepasbaar is. In vele praktische situaties kan de veronderstelling van transitiviteit leiden tot een goede afweging tussen een gereduceerde complexiteit en een bijkomende foutenlast.

Een tweede aspect van de complexiteit van de evaluator is het zoeken naar $\widetilde{p}_{\max}$. Dit heeft een kwadratische complexiteit:

$$O(|A \backslash B||B \backslash A|) \tag{4.30}$$

maar kan worden vereenvoudigd door de possibilistische waarheidswaarden in $\mathbf{M}_{A,B}$ te sorteren. Het sorteren heeft de volgende complexiteit:

$$O(|A \backslash B||B \backslash A| \log(|A \backslash B||B \backslash A|)). \tag{4.31}$$

Daarna kan $\widetilde{p}_{\max}$ echter in logaritmische tijd worden gezocht. Wanneer E_U consistent of transitief is (Hoofdstuk 2), dan geeft dit bijkomende kennis over de mogelijke possibilistische waarheidswaarden in de matrix $\mathbf{M}_{A,B}$. Stel dat twee collecties A en B vergeleken moeten worden en stel dat in de aanpak zoals hier beschreven, E_U een consistente evaluator is. Dit heeft tot gevolg dat beide collecties kunnen worden gepartitioneerd in equivalentieklassen met betrekking tot de mogelijkheid van coreferentie. Anders gezegd, de collectie A kan worden gepartitioneerd zodat voor elk koppel van objecten (a_i, a_i') uit de i^{de} equivalentieklasse geldt:

$$\mu_{E_U(a_i,a_i')}(T) = 1. \tag{4.32}$$

Een dergelijke partitie is minimaal als voor objecten uit twee verschillende equivalentieklassen i en j geldt:

$$\mu_{E_U(a_i,a_j)}(F) = 1. \tag{4.33}$$

Stel dat een dergelijke minimale partitie voor A en B wordt gevonden, dan kunnen de elementen worden gesorteerd zodat elementen binnen een equivalentieklasse in groepen bijeenstaan. Het zoeken naar een maximum kan dan worden versneld door gebruik te maken van de kennis:

$$\begin{array}{ll} & \big(\exists(a,b) \in A \times B : \mu_{E_U(a,b)}(F) = 1\big) \\ \Rightarrow & \big(\forall(a',b') \in K(a) \times K(b) : \mu_{E_U(a',b')}(F) = 1\big) \end{array} \tag{4.34}$$

waarbij $K(u)$ de equivalentieklasse van u voorstelt. Wanneer veel collecties onderling moeten worden vergeleken, dan kunnen alle collecties in lineaire tijd worden herschikt, zodat globaal een complexiteitsvoordeel wordt bereikt.

Een derde en laatste aspect van de complexiteit is het aantal keren dat een maximum moet worden gezocht. Wanneer $\widetilde{p}_{\max}$ op meerdere posities voorkomt, wordt voor elke positie een zoekpad geïnitialiseerd dat overeenkomt met een nog onvolledige afbeelding (Algoritme 4.2). Elk van deze zoekpaden wordt stelselmatig verder onderzocht, tot één van de paden wordt gekozen. In het slechtste geval zijn alle possibilistische waarheidswaarden in de matrix $\mathbf{M}_{A,B}$ gelijk aan elkaar. In dit geval kunnen we het aantal knopen in de zoekboom op niveau k schrijven als:

$$\prod_{i=1}^{k}(|A\backslash B| - i + 1)(|B\backslash A| - i + 1) \tag{4.35}$$

zodat het totaal aantal keer dat een maximum moet worden gezocht gelijk is aan:

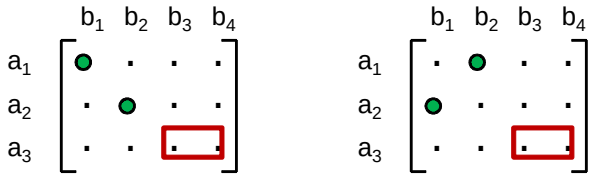
$$\sum_{i=1}^{|A\backslash B|}\left(\prod_{j=1}^{i}(|A\backslash B| - j + 1)(|B\backslash A| - j + 1)\right). \tag{4.36}$$

Zoals is gebleken uit Voorbeeld 4.1 bevat de zoekboom echter heel wat dubbele paden. Dit zijn paden in de zoekboom waarbij dezelfde matrixposities in een andere volgorde zijn geselecteerd. Bovendien is het zo dat de zoekboom ook equivalente paden bevat. Dit zijn paden die overeenkomen met equivalente afbeeldingen.

Definitie 4.6 (Equivalente afbeeldingen)
Twee afbeeldingen ι_1 en ι_2 zijn equivalent als er geldt dat:

$$A_{\iota_1} = A_{\iota_2} \wedge B_{\iota_1} = B_{\iota_2}. \tag{4.37}$$

Paden die tijdens het uitvoeren van Algoritme 4.2 equivalent zijn, geven aanleiding tot eenzelfde oplossing. Hierdoor wordt het aantal te onderzoeken paden gereduceerd. Stel bijvoorbeeld dat twee verzamelingen $A = \{a_1, a_2, a_3\}$ en $B = \{b_1, b_2, b_3, b_4\}$ vergeleken moeten worden. De tijdelijke afbeeldingen $\{(a_1, b_1), (a_2, b_2)\}$ (Figuur 4.3, links) en $\{(a_1, b_2), (a_2, b_1)\}$ (Figuur 4.3, rechts) zijn dan equivalent. De resterende possibilistische waarheidswaarden in de matrix $\mathbf{M}_{A,B}$ zijn voor beide paden dezelfde (rode rechthoek).



Figuur 4.3: Equivalente zoekpaden

Door rekening te houden met equivalente zoekpaden kan het aantal knopen in de zoekboom op niveau k (in het slechtste geval) worden herschreven als:

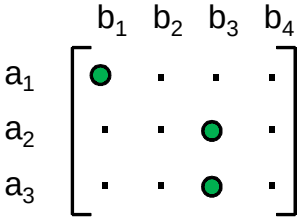
$$\prod_{i=1}^{k} \frac{(|A \backslash B| - i + 1)(|B \backslash A| - i + 1)}{i^2}. \tag{4.38}$$

Dit zorgt voor een verdere reductie van het aantal zoekpaden. Het uitsluiten van paden is echter niet de enige optimalisatie die we kunnen doorvoeren. We beschouwen daarom eerst de volgende definitie.

Definitie 4.7 (Onafhankelijk koppel)
Beschouw twee verzamelingen A en B met $|A| \leq |B|$ en beschouw de matrix $\mathbf{M}_{A,B}$. Wanneer we een leximax-optimale één-op-één afbeelding ι willen construeren op basis van een evaluator E_U, dan noemen we een koppel $(a, b) \in A_\iota \times B_\iota$ een onafhankelijk koppel als:

$$E_U(a, b) = \widetilde{p}_{\max} \tag{4.39}$$
$$\forall b' \in B_\iota \wedge b \neq b' : E_U(a, b') < \widetilde{p}_{\max} \tag{4.40}$$
$$\forall a' \in A_\iota \wedge a \neq a' : E_U(a', b) < \widetilde{p}_{\max}. \tag{4.41}$$

In Figuur 4.4 worden de posities van $\widetilde{p}_{\max}$ aangeduid met een groene schijf. In dit geval is het koppel (a_1, b_1) een onafhankelijk koppel, aangezien er op de rij overeenkomstig met a_1 en op de kolom overeenkomstig met b_1 geen andere positie van $\widetilde{p}_{\max}$ voorkomt.



Figuur 4.4: Onafhankelijke positie (a_1, b_1)

Voor onafhankelijke koppels geldt de volgende eigenschap.

Eigenschap 4.3
Beschouw twee verzamelingen A en B met $|A| \leq |B|$ en beschouw de matrix $\mathbf{M}_{A,B}$. Een onafhankelijk koppel (a, b) behoort steeds tot een leximax-optimale afbeelding.

Bewijs. Beschouw de matrix $\mathbf{M}_{A,B}$ en stel dat $\widetilde{p}_{\max}$ voorkomt op r verschillende rijen en k verschillende kolommen. Dan moeten $\min(r, k)$ koppels (a, b) worden toegevoegd aan ι waarvoor:

$$E_U(a, b) = \widetilde{p}_{\max}. \tag{4.42}$$

Zoniet kan ι geen leximax-optimale afbeelding zijn. Beschouw nu een onafhankelijk koppel (a', b') en stel dat er zou gelden:

$$(a', b') \notin \iota. \tag{4.43}$$

In dit geval kunnen er hoogstens $\min(r, k) - 1$ koppels in ι aanleiding geven tot $\widetilde{p}_{\max}$, wegens de voorwaarden van onafhankelijkheid van (a', b'). Bijgevolg moet een onafhankelijk koppel altijd voorkomen in een leximax-optimale afbeelding. □

Een gevolg van Eigenschap 4.3 is dat voor onafhankelijke koppels geen zoekpad moet worden geïnitialiseerd, waardoor het aantal te onderzoeken paden opnieuw wordt gereduceerd. In sommige gevallen is het niet nodig de afbeelding helemaal te construeren. Wanneer $\widetilde{p}_{\max} = (0, 1)$, zijn alle overgebleven possibilistische waarheidswaarden hieraan gelijk. Alle mogelijke afbeeldingen zijn in dit geval leximax-equivalent. We merken ten slotte op dat een dieptebeperking van de zoekboom kan leiden tot benaderde oplossingen, d.i. afbeeldingen die met grote zekerheid leximax-equivalent zijn. Een dergelijke benadering kan nodig zijn wanneer de gemiddelde kardinaliteit van te vergelijken verzamelingen hoog is.

4.5 Eigenschappen van kwantificatie

Het niet uniek zijn van de leximax-optimale afbeelding ι beperkt de conditionele necessiteit tot het *kardinaliteitsgebaseerde* geval. We zullen in wat volgt een eenvoudige kwantorfunctie voorstellen en we zullen onderzoeken wat de eigenschappen van deze kwantorfunctie zijn. Dit doen we enerzijds met het oog op concrete toepassingen in Hoofdstuk 6. Anderzijds willen we nagaan in welke gevallen transitiviteit van een evaluator $E_{\mathcal{M}(U)}$ geldt.

4.5.1 Enkelvoudige kwantificatie

In dit deel wordt een eenvoudige kwantorfunctie voorgesteld en worden de eigenschappen van de overeenkomstige evaluator $E_{\mathcal{M}(U)}$ onderzocht. De gepa-

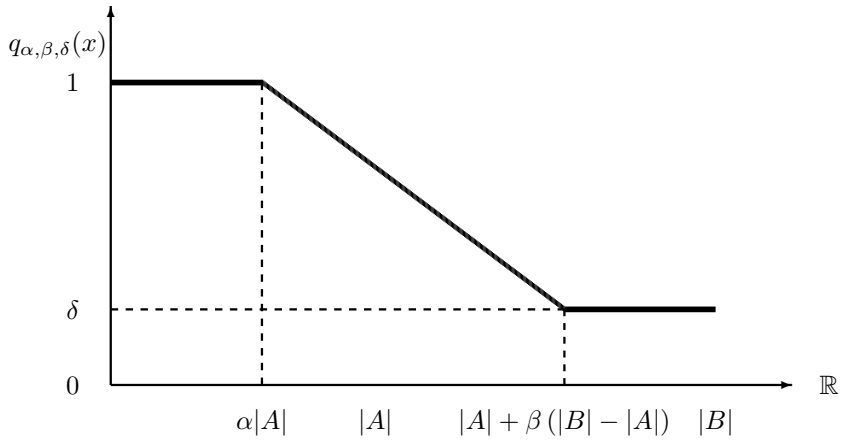
rameteriseerde kwantorfunctie die hier wordt gebruikt is de volgende:

$$q_{\alpha,\beta,\delta}(x) = \begin{cases} 1 & \textbf{als} \quad x < \alpha|A| \\ \delta & \textbf{als} \quad x > |A| + \beta\,(|B| - |A|) \\ 1 - \frac{(1-\delta)(x-\alpha|A|)}{(1-\alpha-\beta)|A|+\beta|B|} & \textbf{anders} \end{cases} \tag{4.44}$$

waarbij $(\alpha, \beta, \delta) \in [0,1]^3$. De vector $\mathbf{w}$, die via genormaliseerde gewichten de conditionele necessiteit voorstelt, kan dan worden berekend als:

$$\forall i \in \{1, ..., |B|\} : \mathbf{w}_i = q_{\alpha,\beta,\delta}(i). \tag{4.45}$$

De kwantorfunctie heeft drie parameters: α en β bepalen de vorm en δ bepaalt een ondergrens voor waarden in $\mathbf{w}$. Een grafische voorstelling van de kwantorfunctie q is te zien in Figuur 4.5.



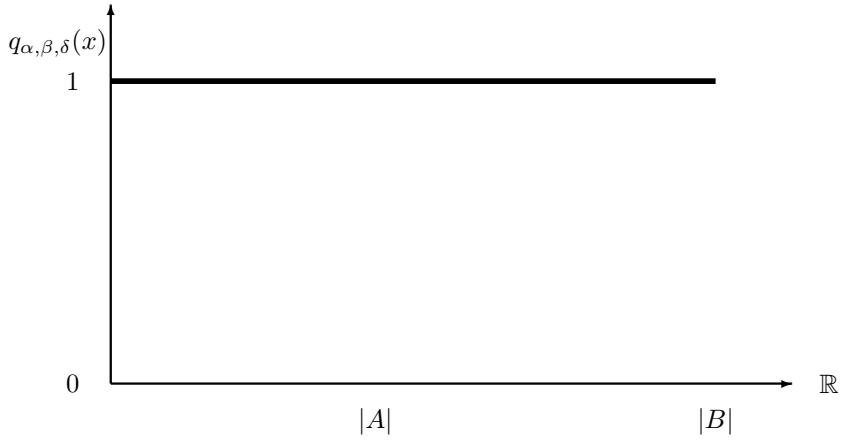
Figuur 4.5: Geparameteriseerde kwantorfunctie voor vergelijking van collecties

De voorgestelde kwantorfunctie is van praktisch belang wegens haar eenvoudige structuur. In Hoofdstuk 6 zal blijken dat een evaluator opvallende eigenschappen vertoont in de context van coreferente karakterstrings als hij gebruik maakt van de hierboven voorgestelde kwantorfunctie. Om die redenen wordt hier een studie gemaakt van de kwantorfunctie en de eigenschappen van de overeenkomstige evaluator.

Laat ons beginnen met het toelichten van twee bijzondere keuzes voor de parameters en de overeenkomstige modellen voor vergelijking van collecties. Als $\delta = 1$ worden de waarden voor parameters α en β irrelevant. De kwantorfunctie is dan een constante functie met ordinaat 1 (Figuur 4.6). Als gevolg zijn alle afgeleide gewichten 1. Dit betekent dat de gegenereerde possibilistische waarheidswaarden na transformatie met T_S ongewijzigd blijven (Hoofdstuk 3). De gebruikte combinatiefunctie is bijgevolg $\widetilde{\wedge}$. Het overeenkomstige model voor vergelijking van collecties vereist dus minstens een strikte gelijkheid van

de kardinaliteiten van de collecties:

$$|A| \neq |B| \Rightarrow E_{\mathcal{M}(U)}(A, B) = (0, 1). \tag{4.46}$$



Figuur 4.6: Een eerste bijzondere kwantorfunctie

Onder de parameterkeuze $\delta = 1$ is $E_{\mathcal{M}(U)}$ sterk reflexief. Bovendien kunnen we nu een interessante stelling in verband met de transitiviteit van $E_{\mathcal{M}(U)}$ bewijzen. Het is namelijk zo dat als $\delta = 1$, transitiviteit van E_U leidt tot transitiviteit van $E_{\mathcal{M}(U)}$. Bovendien is dit de enige keuze voor de parameters waaronder $E_{\mathcal{M}(U)}$ een transitieve evaluator is.

Stelling 4.2 (Transitiviteit van $E_{\mathcal{M}(U)}$)
Een evaluator $E_{\mathcal{M}(U)}$ waarvoor $\delta = 1$ is een transitieve evaluator als en alleen als E_U transitief is.

Bewijs. Het geval waarin de kardinaliteiten van verzamelingen verschillen is triviaal. Het bewijs voor verzamelingen met gelijke kardinaliteiten gebeurt in twee stappen.

(a) Veronderstel drie collecties A, B en C met eenzelfde kardinaliteit en stel dat er geldt:

$$\big(\mathrm{Nec}(p_{(A,B)} = T) > 0\big) \wedge \big(\mathrm{Nec}(p_{(B,C)} = T) > 0\big). \tag{4.47}$$

Beschouw de afbeelding $\iota_{B,C} \circ \iota_{A,B}$ als volgt:

$$\begin{aligned} \forall (u, v) \in U^2 \quad &: \quad (u, v) \in \iota_{B,C} \circ \iota_{A,B} \\ &\Leftrightarrow \quad \exists w \in U : (u, w) \in \iota_{A,B} \wedge (w, v) \in \iota_{B,C}. \end{aligned} \tag{4.48}$$

(a.1) Als zou gelden dat $\iota_{A,C} = \iota_{B,C} \circ \iota_{A,B}$ dan volgt er, gelet op de transitiviteit van E_U, dat:

$$\forall (a,c) \in \iota_{A,C} : \exists (a,b) \in \iota_{A,B} : \mathrm{Nec}(p_{(a,c)} = T) \geq \min \left(\begin{array}{l} \mathrm{Nec}(p_{(a,b)} = T), \\ \mathrm{Nec}(p_{(b,c)} = T) \end{array} \right). \tag{4.49}$$

Hieruit volgt dat:

$$\bigwedge_{(a,c) \in \iota_{A,C}} \mathrm{Nec}(p_{(a,c)} = T) \geq \min \left(\begin{array}{l} \bigwedge_{(a,b) \in \iota_{A,B}} \mathrm{Nec}(p_{(a,b)} = T), \\ \bigwedge_{(b,c) \in \iota_{B,C}} \mathrm{Nec}(p_{(b,c)} = T) \end{array} \right). \tag{4.50}$$

(a.2) In het andere geval is $\iota_{A,C} \neq \iota_{B,C} \circ \iota_{A,B}$, hetgeen betekent dat er minstens twee koppels verschillen tussen de beide afbeeldingen. Er bestaan dan zeker twee koppels (a_1, c_1) en (a_2, c_2) waarvoor er geldt dat:

$$(a_1, c_1) \in \iota_{A,C} \tag{4.51}$$
$$(a_2, c_2) \in \iota_{A,C} \tag{4.52}$$
$$(a_1, c_2) \in \iota_{B,C} \circ \iota_{A,B} \tag{4.53}$$
$$(a_2, c_1) \in \iota_{B,C} \circ \iota_{A,B}. \tag{4.54}$$

Door de leximax-efficiëntie van $\iota_{A,C}$ kan dit verschil worden veroorzaakt door één van twee volgende situaties. In de eerste situatie is:

$$E_U(a_1, c_1) \geq E_U(a_2, c_1) \mathbin{\widetilde{\wedge}} E_U(a_1, c_2). \tag{4.55}$$

Door de transitiviteit van E_U moet er in dit geval gelden dat:

$$\mathrm{Nec}(p_{(a_1,a_2)} = T) \geq \mathrm{Nec}(p_{(a_1,c_1)} = T) \wedge \mathrm{Nec}(p_{(c_1,a_2)} = T) \tag{4.56}$$
$$\mathrm{Nec}(p_{(a_2,c_2)} = T) \geq \mathrm{Nec}(p_{(a_2,a_1)} = T) \wedge \mathrm{Nec}(p_{(a_1,c_2)} = T). \tag{4.57}$$

Substitutie van de eerste in de tweede uitdrukking levert dan:

$$\mathrm{Nec}(p_{(a_2,c_2)} = T) \geq \min \left(\begin{array}{l} \mathrm{Nec}(p_{(a_1,c_1)} = T), \\ \mathrm{Nec}(p_{(c_1,a_2)} = T), \\ \mathrm{Nec}(p_{(a_1,c_2)} = T) \end{array} \right) \tag{4.58}$$

waaruit kan worden besloten dat de kleinste necessiteit voor T onder $\iota_{A,C}$ minstens even groot moet zijn dan de kleinste necessiteit voor T onder $\iota_{B,C} \circ \iota_{A,B}$. In de tweede situatie is:

$$E_U(a_2, c_2) \geq E_U(a_2, c_1) \mathbin{\widetilde{\wedge}} E_U(a_1, c_2). \tag{4.59}$$

Door de transitiviteit van E_U moet in dit geval gelden dat:

$$\mathrm{Nec}(p_{(c_1,c_2)} = T) \geq \mathrm{Nec}(p_{(c_1,a_2)} = T) \wedge \mathrm{Nec}(p_{(a_2,c_2)} = T) \tag{4.60}$$
$$\mathrm{Nec}(p_{(a_1,c_1)} = T) \geq \mathrm{Nec}(p_{(a_1,c_2)} = T) \wedge \mathrm{Nec}(p_{(c_2,c_1)} = T) \tag{4.61}$$

Substitutie van de eerste in de tweede uitdrukking levert dan:

$$\mathrm{Nec}(p_{(a_1,c_1)} = T) \geq \min \left(\begin{array}{l} \mathrm{Nec}(p_{(a_1,c_2)} = T), \\ \mathrm{Nec}(p_{(c_1,a_2)} = T), \\ \mathrm{Nec}(p_{(a_2,c_2)} = T) \end{array} \right). \tag{4.62}$$

Opnieuw kan worden besloten dat de kleinste necessiteit voor T onder $\iota_{A,C}$ minstens even groot moet zijn dan de kleinste necessiteit voor T onder $\iota_{B,C} \circ \iota_{A,B}$. In elk van de bestudeerde gevallen leidt de gestelde voorwaarde tot:

$$\bigwedge_{(a,c)\in\iota_{A,C}} \mathrm{Nec}(p_{(a,c)} = T) \geq \min \left(\begin{array}{l} \bigwedge_{(a,b)\in\iota_{A,B}} \mathrm{Nec}(p_{(a,b)} = T), \\ \bigwedge_{(b,c)\in\iota_{B,C}} \mathrm{Nec}(p_{(b,c)} = T) \end{array} \right) \tag{4.63}$$

waaruit de transitiviteit van $E_{\mathcal{M}(U)}$ volgt in geval (a).

(b) Veronderstel drie collecties A, B en C met dezelfde kardinaliteit en stel dat er geldt:

$$\big(\mathrm{Nec}(p_{(A,B)} = T) > 0\big) \wedge \big(\mathrm{Nec}(p_{(B,C)} = F) > 0\big). \tag{4.64}$$

Dit kunnen we uitdrukken in termen van possibilistische waarheidswaarden als volgt:

$$E_{\mathcal{M}(U)}(A, B) = (1, x) \wedge E_{\mathcal{M}(U)}(B, C) = (y, 1). \tag{4.65}$$

met $x \in [0, 1[$ en $y \in [0, 1[$. Stel dat $|A| = |B| = |C| = n$ en dat:

$$m = \Big|\{(b, c) | (b, c) \in \iota_{B,C} \wedge \mathrm{Nec}(p_{(b,c)} = T) > 0\}\Big| \tag{4.66}$$

dan weten we dat $m < n$ (gegeven) en dat:

$$m = \Big|\{(a, c) | (a, c) \in \iota_{A,C} \wedge \mathrm{Nec}(p_{(a,c)} = T) > 0\}\Big| \tag{4.67}$$

omwille van de transitiviteit van E_U enerzijds en de leximax-efficiëntie van $\iota_{A,C}$ anderzijds. We kunnen dus met zekerheid reeds stellen dat:

$$E_{\mathcal{M}(U)}(A, C) = (z, 1). \tag{4.68}$$

Daarbovenop moet worden aangetoond dat $(x < y) \Rightarrow (y = z)$ en $(x \geq y) \Rightarrow (x \geq z)$. Dit zijn immers de voorwaarden voor transitiviteit (Hoofdstuk 2). Laat ons daarom eerst veronderstellen dat $x < y$ en laat ons twee situaties beschouwen.

(b.1) Indien er geldt dat:

$$\forall (b, c) \in \iota_{B,C} : \mathrm{Nec}(p_{(b,c)} = T) > 0 \Rightarrow (\exists (a, b) \in \iota_{A,B} : (a, c) \in \iota_{A,C}) \tag{4.69}$$

dan wil dit zeggen dat $\iota_{A,C}$ gelijk is aan $\iota_{B,C} \circ \iota_{A,B}$ voor de koppels uit $\iota_{B,C}$ waarvoor zekerheid van coreferentie bestaat. Voor een willekeurig ander koppel (b', c') uit $\iota_{B,C}$ geldt er dat:

$$\mathrm{Nec}(p_{(b',c')} = T) = 0 \tag{4.70}$$

maar ook:

$$E_U(b', c') \geq (y, 1). \tag{4.71}$$

Beschouw voor het koppel (b', c') het element $a' \in A$ zodat $(a', b') \in \iota_{A,B}$, dan moet gelden dat:

$$E_U(a', c') = E_U(b', c') \tag{4.72}$$

wegens de transitiviteit van E_U. Dit is zo aangezien $E_U(a', b')$ een mogelijkheid voor waar heeft die kleiner is of gelijk aan de mogelijkheid voor vals van $E_U(b', c')$. Nu weten we dat (a', c') niet in $\iota_{A,C}$ zit als er een $a'' \in A$ zou bestaan waarvoor:

$$E_U(a'', c') > E_U(a', c'). \tag{4.73}$$

Echter, indien dit zo is, dan heeft de transitiviteit van E_U tot gevolg dat $\iota_{B,C}$ geen leximax-optimale afbeelding is. Bijgevolg is $\iota_{B,C} \circ \iota_{A,B}$ een leximax-optimale afbeelding. Aangezien er geldt dat:

$$E_U(a', c') = E_U(b', c') \tag{4.74}$$

worden de possibilistische waarheidswaarden met mogelijkheid voor waar, kleiner dan of gelijk aan 1 onder $\iota_{B,C}$, ook gegenereerd onder $\iota_{A,C}$. Bijgevolg zijn de kleinste possibilistische waarheidswaarden dezelfde. Er geldt dus dat:

$$\widetilde{\bigwedge_{(a,c)\in\iota_{A,C}}} E_U(a, c) = \widetilde{\bigwedge_{(b,c)\in\iota_{B,C}}} E_U(b, c) \tag{4.75}$$

hetgeen betekent dat:

$$E_{\mathcal{M}(U)}(B, C) = E_{\mathcal{M}(U)}(A, C). \tag{4.76}$$

(b.2) Indien de voorwaarde uit situatie (b.1) niet geldt, dan wil dit zeggen dat een situatie zoals getoond in Figuur 4.7 zich voordoet. Een stippellijn toont koppels waarvoor er zekerheid voor vals is en een volle lijn toont koppels waarvoor er zekerheid voor waar is. In deze situatie geldt er dat:

$$\exists (b, c) \in \iota_{B,C} : \mathrm{Nec}(p_{(b,c)} = T) = 0 \tag{4.77}$$

en dat er een $a \in A$ bestaat zodat $(a, c) \in \iota_{A,C}$ waarbij:

$$\mathrm{Nec}(p_{(a,c)} = T) > 0. \tag{4.78}$$

Dan moet er een $b' \in B$ bestaan waarvoor $(a, b') \in \iota_{A,B}$ zodat:

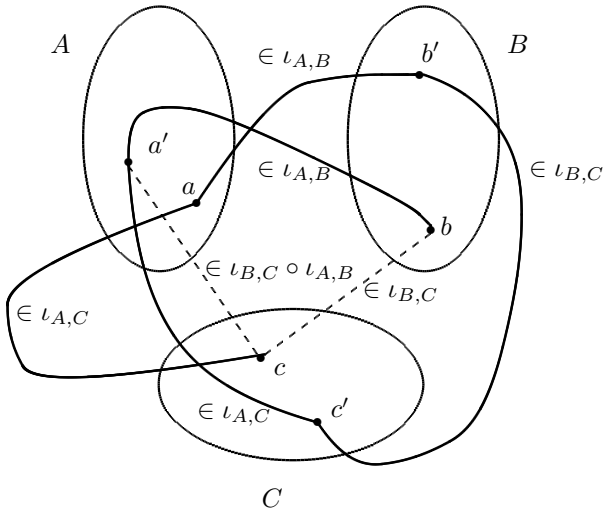
$$\mathrm{Nec}(p_{(a,b')} = T) > 0 \tag{4.79}$$

en dus ook:

$$\mathrm{Nec}(p_{(b',c)} = T) > 0. \tag{4.80}$$

Door de leximax-efficiëntie moet er gelden dat:

$$E_U(a, c) \geq E_U(a, c') \tag{4.81}$$



Figuur 4.7: Transitiviteit

zodat door de transitiviteit van E_U moet gelden:

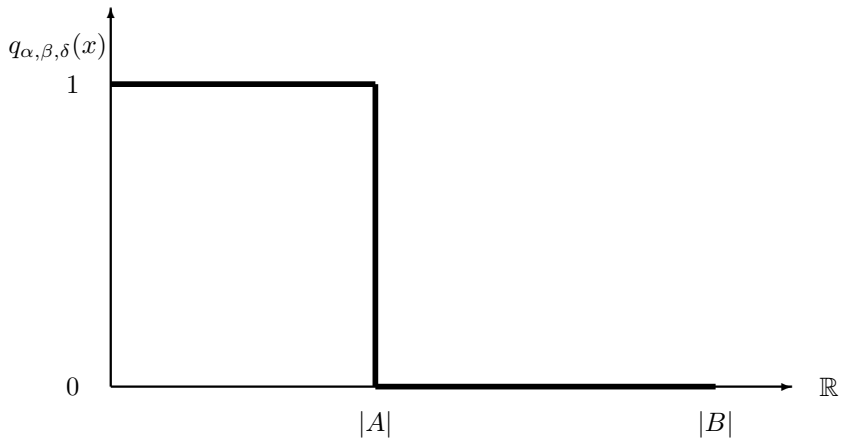
$$E_U(a', c') = E_U(a', c). \tag{4.82}$$

Het gevolg hiervan is dat het afbeelden van a' op c' en a op c geen verschil maakt in de gegenereerde possibilistische waarheidswaarden met mogelijkheid voor vals gelijk aan 1. Anders gezegd, de possibilistische waarheidswaarden met mogelijkheid voor vals gelijk aan 1 zijn dezelfde voor $\iota_{A,C}$ en $\iota_{B,C} \circ \iota_{A,B}$. Opnieuw vinden we dat aan de transitiviteitsvoorwaarden is voldaan. Indien $(x \geq y)$ dan moeten we aantonen dat $x \geq z$. Dit is nu triviaal, aangezien $x < z$ tot gevolg heeft dat $z = y$ wegens hetgeen zonet bewezen is. □

Een tweede bijzonder geval van de kwantorfunctie doet zich voor als de parameters (α, β, δ) gelijk zijn aan $(1, 0, 0)$. In dit geval is de kwantorfunctie een blokfunctie (Figuur 4.8). Onder deze keuze voor de parameters wordt er vereist dat de entiteiten beschreven door A een deelverzameling zijn van de entiteiten beschreven door B. Bijgevolg is $E_{\mathcal{M}(U)}$ niet sterk reflexief.

Wanneer E_U een tweewaardige evaluator is (Hoofdstuk 2), dan maakt onze methode een evaluatie van de doorsnede van twee collecties. Het gebruik van kwantificatie evalueert dan de grootte van de doorsnede, relatief ten opzichte van de kardinaliteiten van de twee collecties die worden vergeleken.

Laat ons nu de meer algemene eigenschappen van kwantificatie onderzoeken. We beginnen met de toekenning van possibiliteit in functie van de gegeven parameters. Er kan worden aangetoond dat puntsgewijs hogere waarden voor de parametervector leiden tot een strengere toekenning van possibiliteit.



Figuur 4.8: Een tweede bijzondere kwantorfunctie

Stelling 4.3
Beschouw een evaluator voor collecties $E_{\mathcal{M}(U)}$, dan wordt $E_{\mathcal{M}(U)}$ strenger naarmate de parametervector (α, β, δ) puntsgewijze groter wordt. Anders gezegd:

$$\begin{aligned}&(\alpha_1 \geq \alpha_2) \wedge (\beta_1 \geq \beta_2) \wedge (\delta_1 \geq \delta_2)\\ &\Rightarrow \left(\forall (A,B) \in \mathcal{M}(U)^2 : E_{\mathcal{M}(U),1}(A,B) \leq E_{\mathcal{M}(U),2}(A,B)\right).\end{aligned} \tag{4.83}$$

waarbij $E_{\mathcal{M}(U),1}$ de parameterwaarden $(\alpha_1, \beta_1, \delta_1)$ aanneemt en $E_{\mathcal{M}(U),2}$ de parameterwaarden $(\alpha_2, \beta_2, \delta_2)$.

Bewijs. Het bewijs van deze stelling is triviaal, aangezien een puntsgewijze grotere parametervector leidt tot een puntsgewijze grotere gewichtsvector, zodat getransformeerde possibilistische waarheidswaarden onder T_S kleiner worden. □

Naast de toekenning van possibiliteiten aan koppels van objecten, zijn we ook geïnteresseerd in de rangorde $\leq_{E_U}$ die door een evaluator E_U wordt gecreëerd in U^2. Meer bepaald noteren we voor twee objectkoppels (u_1, v_1) en (u_2, v_2):

$$(u_1, v_1) \leq_{E_U} (u_2, v_2) \Leftrightarrow E_U(u_1, v_1) \leq E_U(u_2, v_2). \tag{4.84}$$

Hoe groter de zekerheid dat een koppel van objecten coreferent is, hoe hoger dit koppel in de rangorde komt. In het kader van de kwantificatie die hier wordt onderzocht, is het interessant om na te gaan of de rangorde invariant is onder één van de parameters. Meer bepaald willen we weten of er een parameter bestaat waarvoor er geldt dat:

$$\forall (x_1, x_2) \in [0,1]^2 : \leq_{E_{U,1}} \equiv \leq_{E_{U,2}} \tag{4.85}$$

waarbij evaluator $E_{U,1}$ de waarde x_1 aanneemt voor de parameter in kwestie en evaluator $E_{U,2}$ de waarde x_2. Een parameter waarvoor de rangorde invariant is,

speelt immers geen rol in de ordening van koppels van objecten, waardoor hij overbodig wordt. Het kan echter eenvoudig worden ingezien dat in het algemene geval de rangorde niet invariant is ten opzichte van α, β en δ. Desondanks bestaan er enkele speciale gevallen waaraan we aandacht willen besteden. De parameters β en δ spelen bijvoorbeeld geen rol bij de vergelijking van collecties met dezelfde kardinaliteit. In het algemene geval blijkt dat de invloed van β en δ op de rangorde beperkt is. Voor twee collecties A en B geldt er namelijk dat:

$$\begin{aligned} &\exists(\alpha,\beta,\delta) \in [0,1]^3 : \mu_{E_{\mathcal{M}(U)}(A,B)}(T) = 1 \\ &\Rightarrow \forall(\beta,\delta) \in [0,1] \times [0,1[: \mu_{E_{\mathcal{M}(U)}(A,B)}(T) = 1 \end{aligned} \tag{4.86}$$

en

$$\begin{aligned} &\exists(\alpha,\beta,\delta) \in [0,1]^3 : E_{\mathcal{M}(U)}(A,B) = (y,1) \\ &\Rightarrow \forall(\beta,\delta) \in [0,1] \times [0,1[: E_{\mathcal{M}(U)}(A,B) = (y,1). \end{aligned} \tag{4.87}$$

Dit betekent dat de waarheidswaarde die volledig mogelijk wordt geacht, uitsluitend wordt bepaald door α, tenzij $\delta = 1$. Dit is enerzijds te wijten aan de vorm van de transformatiefunctie T_S (Stelling 3.9) en anderzijds aan het feit dat enkel α bepaalt welke gewichten uit de gewichtsvector $\mathbf{w}$ gelijk zijn aan 1. Omwille hiervan zeggen we dat de rangorde $\leq_{E_U}$ gedeeltelijk β-invariant en gedeeltelijk δ-invariant is. Deze gedeeltelijke invarianties spelen een bijzondere rol bij het zoeken naar optimale parameters op basis van een trainingscollectie. Dit wordt behandeld in de context van coreferente karakterstrings (Hoofdstuk 6).

Naast het geval waarin $\delta = 1$ geldt transitiviteit van $E_{\mathcal{M}(U)}$ nooit. Een onderzoek zoals gevoerd in [68], waarbij een meer algemene vorm van transitiviteit (d.i. t-transitiviteit) wordt onderzocht, heeft voor de gebruikte kwantorfunctie geen zin, aangezien er geen graduele compensatie is voor het verschil in kardinaliteit.

4.5.2 Meervoudige kwantificatie

Tot hier toe is een bijzonder eenvoudige kwantorfunctie vooropgesteld. Deze functie heeft een aantal voordelen dat haar gebruik in de context van vergelijking van karakterstrings sterk stimuleren en bovendien leidt dit gebruik tot bijzonder accurate detectie van coreferente strings (Hoofdstuk 6). Echter, de kennis die nodig is om coreferente collecties te identificeren laat zich niet altijd makkelijk voorstellen aan de hand van één enkele kwantorfunctie.

Om dit nadeel aan te pakken bestuderen we meervoudige kwantificatie [76]. In dit raamwerk wordt nog steeds gebruik gemaakt van één kwantorfunctie, die in ons geval wordt gegeven door $q_{\alpha,\beta,\delta}$. De parameters (en dus ook de gegenereerde gewichten) hangen echter af van de te vergelijken collecties A en B. Binnen de context van deze thesis gebruiken we een partitie van het eenheidsinterval met m convexe klassen K_i die allen onderling disjunct zijn:

$$\bigcup_{i=1}^{m} K_i = [0,1] \tag{4.88}$$

Aan elk van deze klassen wordt een parametervector toegekend, die voor klasse K_i gegeven wordt door $(\alpha_i, \beta_i, \delta_i)$. Wanneer twee collecties A en B vergeleken moeten worden met $E_{\mathcal{M}(U)}$, dan wordt de parametervector gekozen die overeenkomt met het interval waarin de kardinaliteitsverhouding van de twee collecties gelegen is. Deze kardinaliteitsverhouding is gelijk aan:

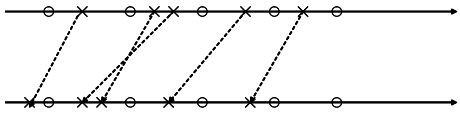
$$\frac{\min(|A|, |B|)}{\max(|A|, |B|)} \tag{4.89}$$

aangenomen dat $\max(|A|, |B|) > 0$. Voor de vergelijking van twee collecties A en B worden dan parameters $(\alpha_i, \beta_i, \delta_i)$ gebruikt als en alleen als:

$$\frac{\min(|A|, |B|)}{\max(|A|, |B|)} \in K_i. \tag{4.90}$$

Een voordeel van meervoudige kwantificatie is dat er een meer flexibele methode voorhanden is om kennis voor te stellen. In woorden kunnen we stellen dat het enkelvoudige model stelt dat twee collecties coreferent zijn als de meeste elementen coreferent zijn, maar waarbij de invulling van de term "meeste" invariant is. Bij het meervoudige model hangt de term "meeste" af van de collecties die worden vergeleken. Typisch zal in een meervoudig model een hoge waarde voor α worden gekozen bij lage kardinaliteitsverhoudingen om dit gedrag in kaart te brengen.

Door het gebruik van meervoudige kwantificatie is het mogelijk om een fijner vergelijkingsmodel te definiëren. Echter, gelet op Stelling 4.3 kan het gebruik van verschillende parametervectoren een voorkeur in de rangorde veroorzaken. Laat ons dit verduidelijken aan de hand van een voorbeeld. Beschouw Figuur 4.9, waarbij tweemaal een rangorde van koppels wordt getoond. We beschouwen een partitie van het eenheidsinterval in twee delen. Een koppel van objecten waarvoor de kardinaliteitsverhouding in het eerste deel ligt wordt aangeduid met $\times$. Een koppel van objecten waarvoor de kardinaliteitsverhouding in het andere deel ligt, wordt aangeduid met $\circ$. De bovenste as toont de rangorde bij het gebruik van enkelvoudige kwantificatie. De onderste as toont de rangorde bij gebruik van meervoudige kwantificatie met $m = 2$. Bij de meervoudige aanpak is voor de kardinaliteitsverhoudingen overeenstemmend met interval $\times$ een puntsgewijze hogere parametervector gekozen. Door Stelling 4.3 weten we dat de toegekende possibilistische waarheidswaarden kleiner zullen zijn. Bijgevolg vindt er een verschuiving van koppels in de rangorde plaats, omdat koppels die worden beoordeeld met minder strengere parameters een voorkeur in de rangorde krijgen. Hoewel het bestaan van deze voorkeur niet noodzakelijk een negatieve impact heeft op $E_{\mathcal{M}(U)}$, is de wetenschap dat deze voorkeur bestaat een belangrijk inzicht in de werking van meervoudige kwantificatie. Het gebruik van meervoudige kwantificatie zal verder worden onderzocht in Hoofdstuk 6.



Figuur 4.9: Enkelvoudige kwantificatie (boven) versus tweevoudige kwantificatie (onder)

4.6 Gedeeltelijke coreferentie

Zoals vermeld in Hoofdstuk 2 is de problematiek van coreferente collecties nauw verwant met het geval waarbij een veralgemeende referentiefunctie wordt gebruikt. Binnen dit hoofdstuk hebben we collecties vergeleken vanuit het standpunt van coreferentie. Dit betekent dat de beschreven entiteiten gelijk moeten zijn. De voorgestelde aanpak laat echter toe om ook gedeeltelijke coreferentie te behandelen, waarbij verondersteld wordt dat objecten een groep van entiteiten beschrijven. Dit kunnen we inzien als volgt. In het geval van $\subseteq$-coreferentie of $\cap$-coreferentie komt de vergelijking van collecties nog steeds neer op het zoeken naar een aantal coreferente elementen in beide collecties. In het geval van $\subseteq$-coreferentie moeten alle entiteiten beschreven door de ene collectie ook beschreven worden door de andere collectie. In het geval van $\cap$-coreferentie mag de doorsnede van beschreven entiteiten niet leeg zijn. Dit betekent dat we nog steeds een één-op-één afbeelding moeten construeren tussen twee collecties. Het verschil tussen coreferentie en gedeeltelijke coreferentie uit zich in de conditionele necessiteit die wordt gebruikt. We hebben hier aangetoond dat deze conditionele necessiteit *kardinaliteitsgebaseerd* moet zijn. Collecties zijn bijgevolg coreferent als de meeste van hun elementen coreferent zijn. Het is precies de invulling van het concept "de meeste" dat verschilt bij gedeeltelijke coreferentie.

4.7 Conclusie

In dit hoofdstuk is onderzocht hoe coreferente collecties van objecten herkend kunnen worden. Hierbij wordt verondersteld dat twee collecties coreferent zijn als hun elementen dezelfde entiteiten beschrijven. De beschreven aanpak voor het vergelijken van collecties vertrekt van een één-op-één afbeelding tussen twee collecties, waarvan de constructie wordt gestuurd door een evaluator op het niveau van elementen. Deze afbeelding wordt zo bepaald dat de koppels ervan aanleiding geven tot een vector van possibilistische waarheidswaarden die leximax-optimaal is. Een dergelijke vector kan worden gecombineerd met de methode voor kenniscombinatie uit Hoofdstuk 3. Aangezien meerdere afbeeldingen leximax-optimaal kunnen zijn, moet de conditionele necessiteit *kardinaliteitsgebaseerd* zijn. Op die manier is de conditionele necessiteit immers onafhankelijk van de gebruikte afbeelding. Hoewel de complexiteit van het zoeken naar een leximax-optimale afbeelding theoretisch hoog kan oplopen,

bestaan er maatregelen om de complexiteit aanzienlijk te verlagen. Een geparameteriseerde kwantorfunctie is voorgesteld en de eigenschappen van deze functie zijn onderzocht. Ten slotte is een model voor meervoudige kwantificatie ingevoerd, waarbij de parameters van de kwantorfunctie afhangen van de collecties die worden vergeleken.

Hoofdstuk 5

Inconsistenties van evaluatoren

5.1 Inleiding

In voorgaande hoofdstukken is bestudeerd hoe de possibiliteitstheorie kan worden gebruikt om de onzekerheid over coreferentie van twee objecten te modelleren. In een praktische toepassing volstaat dit echter niet en is het nodig om concrete beslissingen te nemen. We willen dan een coreferentierelatie $\leftrightarrow$ over het universum O gaan opbouwen op basis van deze concrete beslissingen. De constructie van een dergelijke relatie vormt het onderwerp van dit hoofdstuk. Meer bepaald wordt in dit hoofdstuk bestudeerd hoe een equivalentierelatie over een universum van objecten O kan worden opgebouwd op basis van een evaluator E_O. Een dergelijke relatie induceert steeds een partitie van het universum O. Bij de constructie van een relatie op basis van E_O wordt er rekening gehouden met het feit dat E_O niet noodzakelijk transitief is. Zoals reeds eerder is aangehaald, is het afdwingen van transitiviteit voor een evaluator meestal niet eenvoudig. Niet-transitiviteit wordt veroorzaakt door een gebrek aan kennis, waardoor een evaluator niet in elke situatie correct kan oordelen over coreferentie. Bijgevolg zal een niet-transitieve evaluator leiden tot een aantal inconsistente beslissingen. We onderzoeken hier hoe dergelijke inconsistenties kunnen worden aangepakt. Los daarvan kan er in bepaalde gevallen kennis voorhanden zijn over het maximaal aantal objecten waarmee een gegeven object coreferent is. Bij een klassieke identificatietaak bijvoorbeeld, worden twee collecties van objecten met elkaar vergeleken, waarbij de randvoorwaarde geldt dat elk object maximaal met één ander object coreferent kan zijn.

In dit hoofdstuk vertrekken we van de volgende probleemstelling: gegeven een evaluator E_O, creëer een partitie op O zodat aan gegeven randvoorwaarden is voldaan. Eerst wordt in Sectie 5.2 een beslissingsmodel gedefinieerd als een functie die possibilistische waarheidswaarden afbeeldt op een twee- of

driewaardige beslissingsruimte. Beslissingsmodellen worden gebruikt om een beslissing te nemen over de waarheidswaarde van een Boolese propositie. Een beslissingsmodel in combinatie met een evaluator E_O geeft in de context van coreferentiebepaling aanleiding tot een binaire relatie R over O. In Sectie 5.3 wordt bestudeerd hoe een dergelijke relatie kan worden aangepast zodat een consistente benadering van de coreferentierelatie $\leftrightarrow$ wordt verkregen. In Sectie 5.4 onderzoeken we hoe de ingevoerde aanpak kan worden gebruikt in een context van gedeeltelijke coreferentie. In Sectie 5.5 wordt bestudeerd hoe randvoorwaarden opgelegd kunnen worden en Sectie 5.6 geeft een overzicht van de belangrijkste bevindingen.

5.2 Definities

In deze sectie worden een aantal inleidende definities gegeven. We definiëren eerst een tweewaardig beslissingsmodel.

Definitie 5.1 (Tweewaardig beslissingsmodel)
Gegeven $z \in [0,1]$. Een tweewaardig z-beslissingsmodel is een functie:

$$\mathscr{B} : \mathcal{F}(\mathbb{B}) \to \mathbb{B} \tag{5.1}$$

zodat:

$$\mathscr{B}(\widetilde{p}) = \begin{cases} T & \textbf{\textit{als}} \quad \mathrm{Nec}(p = T) > z \\ F & \textbf{\textit{anders}}. \end{cases} \tag{5.2}$$

De waarde z wordt de drempelwaarde genoemd.

Herinner voor de duidelijkheid uit Hoofdstuk 1 dat er geldt:

$$\mathrm{Nec}(p = T) = 1 - \mathrm{Pos}(p = F) \tag{5.3}$$
$$\mathrm{Nec}(p = F) = 1 - \mathrm{Pos}(p = T). \tag{5.4}$$

Een tweewaardig beslissingsmodel transformeert een possibilistische waarheidswaarde naar een Boolese waarheidswaarde door een ondergrens z te plaatsen op de aanwezige zekerheid voor waar. Een tweewaardig beslissingsmodel modelleert een Boolese beslissing op basis van de kennis beschreven door een possibilistische waarheidswaarde $\widetilde{p}$. Een beslissingsmodel waarvoor $z = 0$ noemen we een maximaal beslissingsmodel en een beslissingsmodel waarvoor $z = 1$ noemen we een minimaal beslissingsmodel. Bij een minimaal beslissingsmodel is de beslissing steeds gelijk aan F. De volgende stelling toont het verband tussen enerzijds de Boolese operatoren $\wedge$ en $\vee$ en anderzijds hun Zadeh-uitbreidingen.

Stelling 5.1
Gegeven een tweewaardig beslissingsmodel $\mathscr{B}$, $\widetilde{p} \in \mathcal{F}(\mathbb{B})$ en $\widetilde{q} \in \mathcal{F}(\mathbb{B})$. Er geldt dat:

$$\mathscr{B}(\widetilde{p}) \wedge \mathscr{B}(\widetilde{q}) = \mathscr{B}(\widetilde{p} \mathbin{\widetilde{\wedge}} \widetilde{q}) \tag{5.5}$$
$$\mathscr{B}(\widetilde{p}) \vee \mathscr{B}(\widetilde{q}) = \mathscr{B}(\widetilde{p} \mathbin{\widetilde{\vee}} \widetilde{q}). \tag{5.6}$$

Bewijs. We bewijzen enkel het geval van $\wedge$. Het geval van $\vee$ is volledig analoog. Enerzijds geldt er dat:

$$\begin{aligned} & \mathscr{B}(\widetilde{p}) \wedge \mathscr{B}(\widetilde{q}) = F && (5.7)\\ \Leftrightarrow \quad & (\mathrm{Nec}(p = T) \leq z) \vee (\mathrm{Nec}(q = T) \leq z) && (5.8)\\ \Leftrightarrow \quad & \min(\mathrm{Nec}(p = T), \mathrm{Nec}(q = T)) \leq z && (5.9)\\ \Leftrightarrow \quad & \mathscr{B}(\widetilde{p} \mathbin{\widetilde{\wedge}} \widetilde{q}) = F && (5.10) \end{aligned}$$

en anderzijds geldt er dat:

$$\begin{aligned} & \mathscr{B}(\widetilde{p}) \wedge \mathscr{B}(\widetilde{q}) = T && (5.11)\\ \Leftrightarrow \quad & (\mathrm{Nec}(p = T) > z) \wedge (\mathrm{Nec}(q = T) > z) && (5.12)\\ \Leftrightarrow \quad & \min(\mathrm{Nec}(p = T), \mathrm{Nec}(q = T)) > z && (5.13)\\ \Leftrightarrow \quad & \mathscr{B}(\widetilde{p} \mathbin{\widetilde{\wedge}} \widetilde{q}) = T. && (5.14) \end{aligned}$$

Hieruit volgt het gestelde in het geval van $\wedge$. □

De keuze voor $\mathbb{B}$ als beslissingsruimte impliceert dat elke possibilistische waarheidswaarde aanleiding geeft tot een eenduidige beslissing (d.i. waar of vals). Het is echter niet ondenkbaar dat het nemen van een beslissing ongewenst is, bijvoorbeeld door een gebrek aan informatie (d.i. een grote onzekerheid). In een dergelijke situatie is een driewaardige beslissingsruimte meer geschikt om beslissingen te modelleren. Om die reden wordt een driewaardig beslissingsmodel gedefinieerd.

Definitie 5.2 (Driewaardig beslissingsmodel)
Gegeven $z_T \in [0,1]$ en $z_F \in [0,1]$. Een driewaardig (z_T, z_F)-beslissingsmodel is een functie:

$$\mathscr{D} : \mathcal{F}(\mathbb{B}) \to \mathcal{P}(\mathbb{B}) \tag{5.15}$$

zodat:

$$\mathscr{D}(\widetilde{p}) = \begin{cases} \{T\} & \textbf{\textit{als}} \quad \mathrm{Nec}(p = T) > z_T \\ \{F\} & \textbf{\textit{als}} \quad \mathrm{Nec}(p = F) > z_F \\ \{T, F\} & \textbf{\textit{anders}}. \end{cases} \tag{5.16}$$

De waarden z_T en z_F worden de drempelwaarden genoemd.

Een driewaardig beslissingsmodel laat toe dat een possibilistische waarheidswaarde teveel onzekerheid bevat om een beslissing te kunnen nemen. In dat geval worden beide waarheidswaarden als mogelijke uitkomst voor een beslissing voorgesteld. In de context van coreferentie kan een driewaardig beslissingsmodel worden gebruikt om aan te geven dat het voor twee objecten o_1 en o_2 niet mogelijk is om een uitspraak te doen over hun coreferentie. Koppels van objecten waarvoor geen uitspraak kan worden gedaan vereisen manuele tussenkomst. In de praktijk wordt een dergelijk mechanisme gebruikt om foute beslissingen van het systeem te minimaliseren, d.i. foute beslissingen worden

vermeden door moeilijke beslissingen niet te nemen en over te laten aan een menselijke gebruiker.

Beslissingsmodellen kunnen nu worden gebruikt om een coreferentierelatie te construeren. In Hoofdstuk 2 is de coreferentierelatie $\leftrightarrow$ gedefinieerd als een equivalentierelatie. Bijgevolg is de relatie $\leftrightarrow$ reflexief, symmetrisch en transitief. We onderzoeken nu hoe een relatie R kan worden afgeleid, zodat deze relatie als benadering van de relatie $\leftrightarrow$ kan worden gebruikt.

Definitie 5.3
Een tweewaardig beslissingsmodel $\mathscr{B}$ geeft aanleiding tot de volgende binaire relaties:

$$R^T_{\mathscr{B}} = \left\{(o_1,o_2)|(o_1,o_2) \in O^2 \wedge \mathscr{B}(E_O(o_1,o_2)) = T\right\} \tag{5.17}$$
$$R^F_{\mathscr{B}} = \left\{(o_1,o_2)|(o_1,o_2) \in O^2 \wedge \mathscr{B}(E_O(o_1,o_2)) = F\right\}. \tag{5.18}$$

Voor een tweewaardig beslissingsmodel geldt er bij definitie dat:

$$R^F_{\mathscr{B}} = \overline{R^T_{\mathscr{B}}}. \tag{5.19}$$

Daarom zullen we in de context van tweewaardige beslissingsmodellen $R^T_{\mathscr{B}}$ noteren als $R_{\mathscr{B}}$.

Stelling 5.2 (Inclusie)
Voor twee tweewaardige beslissingsmodellen $\mathscr{B}_1$ en $\mathscr{B}_2$ met respectievelijke drempelwaarden z_1 en z_2 geldt er dat:

$$(z_2 \leq z_1) \Rightarrow (R_{\mathscr{B}_1} \subseteq R_{\mathscr{B}_2}). \tag{5.20}$$

Bewijs. Voor een willekeurig koppel $(o_1,o_2) \in O^2$ weten we:

$$\widetilde{p}_{(o_1,o_2)} = E_O(o_1,o_2). \tag{5.21}$$

We vinden nu dat:

$$\mathscr{B}_1(\widetilde{p}_{(o_1,o_2)}) = T \tag{5.22}$$
$$\Leftrightarrow \quad \mathrm{Nec}(p_{(o_1,o_2)} = T) > z_1 \tag{5.23}$$
$$\Leftrightarrow \quad \mathrm{Nec}(p_{(o_1,o_2)} = T) > z_2 \tag{5.24}$$
$$\Leftrightarrow \quad \mathscr{B}_2(\widetilde{p}_{(o_1,o_2)}) = T. \tag{5.25}$$
$$\tag{5.26}$$

Bijgevolg geldt er dat:

$$\forall (o_1,o_2) \in O^2 : (o_1,o_2) \in R_{\mathscr{B}_1} \Rightarrow (o_1,o_2) \in R_{\mathscr{B}_2} \tag{5.27}$$

hetgeen betekent dat:

$$R_{\mathscr{B}_1} \subseteq R_{\mathscr{B}_2}. \tag{5.28}$$

□

Door de symmetrie van evaluatoren is $R_{\mathscr{B}}$ steeds een symmetrische relatie en onder de voorwaarde dat $z < 1$ is $R_{\mathscr{B}}$ reflexief. Als we $R_{\mathscr{B}}$ als benadering voor de relatie $\leftrightarrow$ willen gebruiken, dan moet $R_{\mathscr{B}}$ ook transitief zijn. In verband hiermee vinden we het volgende resultaat.

Stelling 5.3
Als E_O een transitieve evaluator is, dan is $R_{\mathscr{B}}$ een transitieve relatie.

Bewijs. Stel dat $\mathscr{B}$ een tweewaardig z-beslissingsmodel is. Voor drie willekeurige objecten o_1, o_2 en o_3 moet er, aangezien E_O een transitieve evaluator is, gelden dat:

$$\left(\wedge \begin{array}{l} \big(\mathrm{Nec}(p_{(o_1,o_2)} = T) \geq z\big) \\ \big(\mathrm{Nec}(p_{(o_2,o_3)} = T) \geq z\big) \end{array} \right) \Rightarrow \big(\mathrm{Nec}(p_{(o_1,o_3)} = T) \geq z\big) \tag{5.29}$$

en

$$\left(\wedge \begin{array}{l} \big(\mathrm{Nec}(p_{(o_1,o_2)} = T) < z\big) \\ \big(\mathrm{Nec}(p_{(o_2,o_3)} = T) < z\big) \end{array} \right) \Rightarrow \big(\mathrm{Nec}(p_{(o_1,o_3)} = T) < z\big)\,. \tag{5.30}$$

Hieruit volgt onmiddellijk dat $R_{\mathscr{B}}$ een transitieve relatie is. □

Stelling 5.3 verklaart waarom een ondergrens wordt geplaatst op de zekerheid voor waar. Het kan namelijk worden ingezien dat het plaatsen van een bovengrens op de zekerheid van F niet zou leiden tot een transitieve relatie als E_O transitief is. Dit komt omdat het complement van een transitieve relatie niet noodzakelijk transitief is. Eens een equivalentierelatie over O beschikbaar is, kan deze relatie worden gebruikt voor het construeren van een partitie van O. Objecten die in dezelfde partitieklasse zitten, worden hierbij beschouwd als coreferent. Een dergelijke partitie vormt een oplossing voor een praktisch coreferentieprobleem.

Definitie 5.4
Als $R_{\mathscr{B}}$ een equivalentierelatie is, dan bepaalt ze een partitie $\mathscr{P}_{R_{\mathscr{B}}} = \{\mathscr{P}_1, \mathscr{P}_2, ...\}$ van O waarvoor er geldt dat:

$$\forall (o_1, o_2) \in \mathscr{P}_i \times \mathscr{P}_i : (o_1, o_2) \in R_{\mathscr{B}}. \tag{5.31}$$

Als elke klasse $\mathscr{P}_i$ een maximale kardinaliteit heeft, dan geldt er dat:

$$\forall i \neq j : \forall (o_1, o_2) \in \mathscr{P}_i \times \mathscr{P}_j : (o_1, o_2) \notin R_{\mathscr{B}} \tag{5.32}$$

en bevat de partitie een minimaal aantal klassen. We spreken dan van een minimale partitie en deze wordt genoteerd als $\mathscr{P}^*_{R_{\mathscr{B}}}$. Een gelijkaardige redenering kunnen we volgen voor driewaardige beslissingsmodellen.

Definitie 5.5
Een driewaardig beslissingsmodel $\mathscr{D}$ geeft aanleiding tot volgende binaire relaties:

$$R^T_{\mathscr{D}} = \big\{(o_1, o_2) | (o_1, o_2) \in O^2 \wedge \mathscr{D}(E_O(o_1, o_2)) = \{T\}\big\} \tag{5.33}$$

$$R^F_{\mathscr{D}} = \big\{(o_1, o_2) | (o_1, o_2) \in O^2 \wedge \mathscr{D}(E_O(o_1, o_2)) = \{F\}\big\} \tag{5.34}$$

$$R^{T,F}_{\mathscr{D}} = \big\{(o_1, o_2) | (o_1, o_2) \in O^2 \wedge \mathscr{D}(E_O(o_1, o_2)) = \{T, F\}\big\}\,. \tag{5.35}$$

Voor deze relaties geldt er bij definitie dat:

$$\overline{R^T_{\mathscr{D}}} = R^F_{\mathscr{D}} \cup R^{T,F}_{\mathscr{D}} \tag{5.36}$$

$$\overline{R^F_{\mathscr{D}}} = R^T_{\mathscr{D}} \cup R^{T,F}_{\mathscr{D}} \tag{5.37}$$

$$\overline{R^{T,F}_{\mathscr{D}}} = R^T_{\mathscr{D}} \cup R^F_{\mathscr{D}}. \tag{5.38}$$

Dit betekent dat er voor een driewaardig beslissingsmodel $\mathscr{D}$ met drempelwaarden z_T en z_F en een tweewaardig beslissingsmodel $\mathscr{B}$ met drempelwaarde $z = z_T$ geldt:

$$R^T_{\mathscr{D}} = R_{\mathscr{B}}. \tag{5.39}$$

Bijgevolg geldt Stelling 5.3 ook voor de relatie $R^T_{\mathscr{D}}$ verkregen uit een driewaardig beslissingsmodel $\mathscr{D}$. Vanuit een logisch standpunt is de transitiviteit van $R^T_{\mathscr{D}}$ niet voldoende om consistente beslissingen te modelleren. Voor drie willekeurige objecten o_1, o_2 en o_3 moet er ook gelden dat:

$$\Big((o_1, o_2) \in R^{T,F}_{\mathscr{D}} \wedge (o_2, o_3) \in R^T_{\mathscr{D}}\Big) \Rightarrow (o_1, o_3) \in R^{T,F}_{\mathscr{D}} \tag{5.40}$$

$$\Big((o_1, o_2) \in R^{T,F}_{\mathscr{D}} \wedge (o_2, o_3) \in R^F_{\mathscr{D}}\Big) \Rightarrow (o_1, o_3) \notin R^T_{\mathscr{D}}. \tag{5.41}$$

In sommige gevallen is aan (5.40) en (5.41) automatisch voldaan, zoals blijkt uit volgende stelling.

Stelling 5.4
Voor een driewaardig beslissingsmodel $\mathscr{D}$ met gelijke drempelwaarden (d.i. $z_T = z_F$) en een transitieve evaluator E_O is er voldaan aan (5.40) *en* (5.41).

Bewijs. We bewijzen eerst dat (5.40) geldt. Gelet op $(o_2, o_3) \in R^T_{\mathscr{D}}$ geldt er:

$$E_O(o_2, o_3) = (1, a) \tag{5.42}$$

met $1 - a > z_T$. Voor (o_1, o_2) kunnen er zich twee gevallen voordoen.
(1) Als er geldt dat:

$$E_O(o_1, o_2) = (1, b) \tag{5.43}$$

dan moet er, gelet op $(o_1, o_2) \in R^{T,F}_{\mathscr{D}}$ gelden dat:

$$1 - b \leq z_T. \tag{5.44}$$

Gelet op de transitiviteit van E_O en het feit dat $1 - b \leq z_T < 1 - a$ geldt er dat:

$$E_O(o_1, o_3) = (1, b) \tag{5.45}$$

waaruit volgt dat:

$$(o_1, o_3) \in R^{T,F}_{\mathscr{D}}. \tag{5.46}$$

(2) Als er geldt dat:

$$E_O(o_1, o_2) = (b, 1) \tag{5.47}$$

dan moet er, gelet op $(o_1, o_2) \in R_{\mathscr{D}}^{T,F}$ gelden dat:

$$1 - b \leq z_F. \tag{5.48}$$

Vermits $1 - a > z_T = z_F \geq 1 - b$ geldt er dat:

$$a < b. \tag{5.49}$$

Gelet op de transitiviteit van E_O geldt er dat:

$$E_O(o_1, o_3) = (b, 1) \tag{5.50}$$

waaruit volgt dat:

$$(o_1, o_3) \in R_{\mathscr{D}}^{T,F}. \tag{5.51}$$

Vervolgens bewijzen we dat (5.41) geldt. Gelet op $(o_2, o_3) \in R_{\mathscr{D}}^{F}$ geldt er:

$$E_O(o_2, o_3) = (a, 1) \tag{5.52}$$

met $1 - a > z_F$. Voor (o_1, o_2) kunnen er zich twee gevallen voordoen.
(1) Als er geldt dat:

$$E_O(o_1, o_2) = (1, b) \tag{5.53}$$

dan moet er, gelet op de transitiviteit van E_O gelden dat:

$$E_O(o_1, o_3) = (c, 1) \tag{5.54}$$

waaruit volgt dat:

$$\mathrm{Nec}(p_{(o_1,o_3)} = T) = 0. \tag{5.55}$$

Bijgevolg geldt er dat:

$$(o_1, o_3) \notin R_{\mathscr{D}}^{T}. \tag{5.56}$$

(2) Als er geldt dat:

$$E_O(o_1, o_2) = (b, 1) \tag{5.57}$$

dan moet er, gelet op $(o_1, o_2) \in R_{\mathscr{D}}^{T,F}$ gelden dat:

$$1 - b \leq z_F. \tag{5.58}$$

Bijgevolg geldt er dat $a \neq b$. Enerzijds, als er geldt dat:

$$E_O(o_1, o_3) = (c, 1) \tag{5.59}$$

dan volgt daaruit onmiddellijk dat:

$$(o_1, o_3) \notin R_{\mathscr{D}}^{T}. \tag{5.60}$$

Anderzijds, als er geldt dat:

$$E_O(o_1, o_3) = (1, c) \tag{5.61}$$

dan moet er ook gelden dat:

$$c \geq \max(a, b) \tag{5.62}$$

waaruit volgt dat:

$$1 - c \leq \min(1 - a, 1 - b) \leq z_F = z_T. \tag{5.63}$$

Dit betekent dat er geldt:

$$(o_1, o_3) \notin R^T_{\mathscr{D}}. \tag{5.64}$$

□

5.3 Constructie van consistente relaties

We hebben in de vorige sectie aangetoond hoe een evaluator kan worden gebruikt om tot een concrete oplossing te komen voor het coreferentieprobleem. In deze sectie willen we onderzoeken hoe inconsistenties in een dergelijke oplossing kunnen worden vermeden. Als E_O bijvoorbeeld geen transitieve evaluator is, dan is $R_{\mathscr{B}}$ niet noodzakelijk een transitieve relatie, waardoor ze niet kan worden gebruikt als benadering voor $\leftrightarrow$. We zullen daarom bespreken hoe een transitieve relatie kan worden afgeleid uit $R_{\mathscr{B}}$ waarbij de kennis van E_O zo goed mogelijk gerespecteerd wordt. In het geval van een driewaardig beslissingsmodel kan dit principe worden toegepast op $R^T_{\mathscr{D}}$. Daarnaast zullen we in het geval van een driewaardig beslissingsmodel aantonen hoe uit $\overline{R^T_{\mathscr{D}}}$ een relatie $R^{T,F}_{\mathscr{D}}$ kan worden geëxtraheerd die voldoet aan (5.40) en (5.41).

5.3.1 Consistentie van $R_{\mathscr{B}}$

Voor een niet-transitieve evaluator is $R_{\mathscr{B}}$ niet noodzakelijk een transitieve relatie. Dat wil zeggen dat volgende situatie zich minstens één keer voordoet:

$$((o_1, o_2) \in R_{\mathscr{B}}) \wedge ((o_2, o_3) \in R_{\mathscr{B}}) \wedge ((o_1, o_3) \notin R_{\mathscr{B}}). \tag{5.65}$$

De evaluator E_O maakt bijgevolg voor minstens één koppel een foute toekenning van possibiliteit en deze fout komt voort uit een inconsistentie in de kennis van E_O. Het is hierbij niet beslist voor welk van de drie koppels een foute toekenning van possibiliteit is gemaakt. Een mogelijke oplossing voor dit probleem is het zoeken naar de transitieve sluiting [77, 78, 79] van $R_{\mathscr{B}}$. De transitieve sluiting van een relatie R is de kleinste transitieve relatie R^+ zodat $R \subseteq R^+$. Als R een transitieve relatie is, dan geldt $R = R^+$. Het vervangen van $R_{\mathscr{B}}$ door zijn transitieve sluiting impliceert dat de inconsistenties van een evaluator worden opgelost door koppels met een tekort aan zekerheid voor waar toch te beschouwen als coreferente koppels. In het geval van (5.65) wil dit zeggen dat o_1 en o_3 als coreferent worden beschouwd. Het kan echter in vraag worden gesteld of het correct is dat een evaluator steeds de zekerheid voor waar onderschat. Een tweede type fout zou kunnen zijn dat een evaluator voor bepaalde

koppels de zekerheid voor waar overschat. In het geval van (5.65) zouden we dan bijvoorbeeld het koppel (o_2, o_3) uit de relatie kunnen halen, hetgeen eveneens leidt tot een transitieve relatie.

Algoritme 5.1 $\mathcal{H}_f(O, \mathscr{B})$

1: $\forall o_i \in O : \mathscr{P}_i \leftarrow \{o_i\}$
2: $\mathscr{P} \leftarrow \{\mathscr{P}_i | o_i \in O\}$
3: continue$\leftarrow$ **true**
4: **while** continue **do**
5: $(k, l) \leftarrow \arg\max_{(a,b)} f(\mathscr{P}_a, \mathscr{P}_b)$
6: **if** $\mathscr{B}(f(\mathscr{P}_k, \mathscr{P}_l))$ **then**
7: $\mathscr{P}_k \leftarrow \mathscr{P}_k \cup \mathscr{P}_l$
8: $\mathscr{P} \leftarrow \mathscr{P} \backslash \{\mathscr{P}_l\}$
9: **else**
10: continue$\leftarrow$ **false**
11: **end if**
12: **end while**
13: **return** $\mathscr{P}$

Om dit idee uit te werken maken we een zijsprong naar de *datamining* en meer bepaald naar het principe van hiërarchisch clusteren (Bijlage A). Deze techniek construeert een partitie over een gegeven datacollectie door stelselmatig twee partitieklassen samen te voegen. Door deze manier van werken wordt een boomstructuur opgebouwd waarbij elke knoop overeen komt met één enkele partitie. Dit verklaart de naam hiërarchisch clusteren. Algoritme 5.1 schetst de aanpak in pseudocode, waarbij we het klassieke algoritme omgevormd hebben naar de possibilistische context van deze thesis. Aanvankelijk wordt elk object o_i in een afzonderlijke partitieklasse geplaatst (regels 1 en 2). Vervolgens worden er iteratief twee partitieklassen gekozen volgens een keuzecriterium (regel 5). Er wordt gecontroleerd of deze partitieklassen mogen worden samengevoegd (regel 6) en zoja, dan worden ze samengevoegd (regels 7 en 8). Op deze manier wordt rechtstreeks een partitie van O verkregen, zonder de relatie $R_{\mathscr{B}}$ expliciet in rekening te brengen. Met de verkregen partitie komt een equivalentierelatie R' overeen die kan worden gezien als een afgeleide relatie van $R_{\mathscr{B}}$. Deze afleiding gebeurt echter niet expliciet bij de berekening van de partitie door Algoritme 5.1. In het algoritme is f een functie:

$$f : \mathcal{P}(O)^2 \rightarrow \mathcal{F}(\mathbb{B}). \tag{5.66}$$

Elke functie f correspondeert met een regel voor het samenvoegen van twee partitieklassen. In het kader van de possibilistische aanpak die we hier handhaven beschouwen we twee bijzondere gevallen, namelijk de existentiële regel:

$$f_{\tilde{\vee}}(\mathscr{P}_i, \mathscr{P}_j) = \widetilde{\bigvee_{(o_1,o_2)\in\mathscr{P}_i\times\mathscr{P}_j}} (E_O(o_1, o_2)) \tag{5.67}$$

en de universele regel:

$$f_{\widetilde{\wedge}}\left(\mathscr{P}_i, \mathscr{P}_j\right) = \widetilde{\bigwedge_{(o_1,o_2)\in\mathscr{P}_i\times\mathscr{P}_j}} \left(E_O(o_1,o_2)\right). \tag{5.68}$$

Voor de existentiële regel zien we dat:

$$\mathscr{B}(f_{\widetilde{\vee}}\left(\mathscr{P}_i, \mathscr{P}_j\right)) = \bigvee_{(o_1,o_2)\in\mathscr{P}_i\times\mathscr{P}_j} \mathscr{B}\left(E_O(o_1,o_2)\right). \tag{5.69}$$

Onder deze regel mogen twee partitieklassen worden samengevoegd als er minstens één koppel van objecten (o_1, o_2) bestaat waarvoor E_O voldoende zekerheid op coreferentie postuleert volgens $\mathscr{B}$. Voor de universele regel zien we dat:

$$\mathscr{B}(f_{\widetilde{\wedge}}\left(\mathscr{P}_i, \mathscr{P}_j\right)) = \bigwedge_{(o_1,o_2)\in\mathscr{P}_i\times\mathscr{P}_j} \mathscr{B}\left(E_O(o_1,o_2)\right). \tag{5.70}$$

Onder deze regel mogen twee partitieklassen worden samengevoegd als E_O voor alle koppels van objecten voldoende zekerheid op coreferentie postuleert volgens $\mathscr{B}$. Het algoritme zoals hier gedefinieerd bezit enkele interessante eigenschappen.

Stelling 5.5
Voor een tweewaardig beslissingsmodel $\mathscr{B}$ geldt er dat:

$$\mathcal{H}_{f_{\widetilde{\vee}}}\left(O, \mathscr{B}\right) = \mathscr{P}^*_{R^+_{\mathscr{B}}}. \tag{5.71}$$

Bewijs. Veronderstel dat $\mathcal{H}_{f_{\widetilde{\vee}}}\left(O, \mathscr{B}\right) = \mathscr{P}$, waarbij $\mathscr{P}$ een partitie is die overeenkomt met een equivalentierelatie R' (Definitie 5.4). Voor de relatie $R_{\mathscr{B}}$ moet er gelden dat:

$$\forall(o_1,o_2) \in R_{\mathscr{B}} : \mathscr{B}(E_O(o_1,o_2)) = T. \tag{5.72}$$

Gelet op de conditie in regel 6 van het algoritme moet er dan gelden:

$$\forall(o_1,o_2) \in R_{\mathscr{B}} : \exists\mathscr{P}_i \in \mathscr{P} : o_1 \in \mathscr{P}_i \wedge o_2 \in \mathscr{P}_i. \tag{5.73}$$

Dit wil zeggen dat als twee objecten gerelateerd zijn door $R_{\mathscr{B}}$, dan moeten deze objecten zeker samen in een partitieklasse zitten. Hieruit volgt dat er geldt:

$$R_{\mathscr{B}} \subset R'. \tag{5.74}$$

Aangezien er nu geldt dat:

$$\forall\mathscr{P}_k \in \mathscr{P} : \forall\mathscr{P}_l \in \mathscr{P}\backslash\{\mathscr{P}_k\} : \forall(o_1,o_2) \in \mathscr{P}_k \times \mathscr{P}_l : (o_1,o_2) \notin R_{\mathscr{B}} \tag{5.75}$$

weten we dat $\mathscr{P}$ een minimale partitie is. Bovendien geldt er voor een willekeurige klasse $\mathscr{P}_k \in \mathscr{P}$ dat als deze klasse wordt opgesplitst in twee disjuncte klassen $\mathscr{P}_{k,1}$ en $\mathscr{P}_{k,2}$ er geldt:

$$\exists(o_1,o_2) \in \mathscr{P}_{k,1} \times \mathscr{P}_{k,2} : (o_1,o_2) \in R_{\mathscr{B}}. \tag{5.76}$$

Dit wil zeggen dat elke transitieve deelrelatie van R' geen *superrelatie* van $R_{\mathscr{B}}$ kan zijn. Bijgevolg is R' de kleinst mogelijke transitieve *superrelatie* van $R_{\mathscr{B}}$. Nog anders gezegd betekent dit dat $R' = R^+_{\mathscr{B}}$. □

Stelling 5.5 stelt dat partitioneren op basis van de existentiële regel leidt tot een minimale partitie op basis van de transitieve sluiting van $R_{\mathscr{B}}$. De transitieve sluiting van $R_{\mathscr{B}}$ geeft met andere woorden aanleiding tot een partitie die door Algoritme 5.1 onder voorwaarden ook kan worden bereikt. Als $f_{\tilde{\vee}}$ aanleiding geeft tot een partitie die voortkomt uit een transitieve sluiting van $R_{\mathscr{B}}$, dan stelt zich onmiddellijk de vraag tot welke partitie $f_{\tilde{\wedge}}$ leidt. We kunnen afleiden dat:

$$\mathcal{H}_{f_{\tilde{\wedge}}}(O, \mathscr{B}) = \mathscr{P}^*_{R'_{\mathscr{B}}} \tag{5.77}$$

waarbij $R'_{\mathscr{B}} \subset R_{\mathscr{B}}$. Interessant genoeg is $R'_{\mathscr{B}}$ niet noodzakelijk een transitieve opening van $R_{\mathscr{B}}$. Dit zal later worden aangetoond met een voorbeeld. Een transitieve opening van een relatie R is een maximale transitieve relatie die een deelrelatie is van R. In het geval van (5.65) worden inconsistenties door het gebruik van $f_{\tilde{\wedge}}$ opgelost door ofwel (o_1, o_2) ofwel (o_2, o_3) uit de relatie te halen. Merk op dat als er voor drie objecten (o_1, o_2, o_3) geldt dat:

$$E_O(o_1, o_2) = E_O(o_2, o_3) \tag{5.78}$$

de keuze willekeurig is. Hierdoor is het resultaat van $\mathcal{H}_{f_{\tilde{\wedge}}}(O, \mathscr{B})$ niet uniek bepaald. Een belangrijk resultaat is de volgende stelling.

Stelling 5.6
Voor een universum van objecten O en een evaluator E_O geldt er dat:

$$\forall \mathscr{P}_i \in \mathcal{H}_{f_{\tilde{\wedge}}}(O, \mathscr{B}) : \exists \mathscr{P}_j \in \mathcal{H}_{f_{\tilde{\vee}}}(O, \mathscr{B}) : \mathscr{P}_i \subseteq \mathscr{P}_j \tag{5.79}$$

Bewijs. Beschouw een willekeurige klasse $\mathscr{P}_i \in \mathcal{H}_{f_{\tilde{\wedge}}}(O, \mathscr{B})$. Voor deze klasse moet er gelden dat:

$$\forall (o_1, o_2) \in \mathscr{P}_i^2 : \mathscr{B}(E_O(o_1, o_2)) = T. \tag{5.80}$$

Dit is een gevolg van het gebruik van de functie $f_{\tilde{\wedge}}$. Dit betekent dat elk koppel van objecten (o_1, o_2) die samen in een klasse zitten na toepassing van $\mathcal{H}_{f_{\tilde{\wedge}}}$, ook noodzakelijk samen in een klasse zitten na toepassing van $\mathcal{H}_{f_{\tilde{\vee}}}$. Elke klasse in de partitie gevormd door $\mathcal{H}_{f_{\tilde{\wedge}}}$ moet bijgevolg een deelverzameling zijn van een klasse in de partitie gevormd door $\mathcal{H}_{f_{\tilde{\vee}}}$. □

Deze belangrijke stelling impliceert dat het resultaat van $\mathcal{H}_{f_{\tilde{\wedge}}}(O, \mathscr{B})$ kan worden berekend door de klassen van $\mathcal{H}_{f_{\tilde{\vee}}}(O, \mathscr{B})$ verder op te splitsen. Meer bepaald geldt er dat:

$$\mathcal{H}_{f_{\tilde{\wedge}}}(O, \mathscr{B}) = \bigcup_{\mathscr{P}_i \in \mathcal{H}_{f_{\tilde{\vee}}}(O, \mathscr{B})} \left(\mathcal{H}_{f_{\tilde{\wedge}}}(\mathscr{P}_i, \mathscr{B}) \right) \tag{5.81}$$

Deze vaststelling is belangrijk aangezien $\mathcal{H}_{f_{\tilde{\vee}}}(O, \mathscr{B})$ efficiënt kan worden berekend. De partitionering op basis van $f_{\tilde{\vee}}$ kan namelijk worden berekend door eerst de relatie $R_{\mathscr{B}}$ te bepalen en vervolgens koppels in de relatie te zoeken waarvan de objecten nog niet in dezelfde klasse zitten. Dit zoeken kan efficiënt gebeuren door eerst alle koppels $(o_1, o_2) \in R_{\mathscr{B}}$ te sorteren volgens $E_O(o_1, o_2)$ en vervolgens de gesorteerde lijst in lineaire tijd te overlopen. De efficiënte berekening van $\mathcal{H}_{f_{\tilde{\vee}}}$ heeft dan als gevolg dat we ook $\mathcal{H}_{f_{\tilde{\wedge}}}$ efficiënt kunnen berekenen door eerst $\mathcal{H}_{f_{\tilde{\vee}}}$ te berekenen en daarna de verkregen klassen verder op te splitsen. Met betrekking tot verschillende beslissingsmodellen kunnen we het volgende inzien.

Stelling 5.7
Gegeven een universum van objecten O en een evaluator E_O. Beschouw twee beslissingsmodellen $\mathscr{B}_1$ en $\mathscr{B}_2$ zodat $z_1 \leq z_2$. Dan geldt er:

$$\forall \mathscr{P}_i \in \mathcal{H}_{f_{\tilde{\vee}}}(O, \mathscr{B}_2) : \exists \mathscr{P}_j \in \mathcal{H}_{f_{\tilde{\vee}}}(O, \mathscr{B}_1) : \mathscr{P}_i \subseteq \mathscr{P}_j. \tag{5.82}$$

Bewijs. Aangezien $z_1 \leq z_2$ geldt er voor twee willekeurige klassen van objecten $\mathscr{P}_k$ en $\mathscr{P}_l$ dat:

$$\mathscr{B}_1\left(f_{\tilde{\vee}}\left(\mathscr{P}_k, \mathscr{P}_l\right)\right) \Rightarrow \mathscr{B}_2\left(f_{\tilde{\vee}}\left(\mathscr{P}_k, \mathscr{P}_l\right)\right). \tag{5.83}$$

□

Tot hier toe hebben we de functies $f_{\tilde{\vee}}$ en $f_{\tilde{\wedge}}$ beschouwd. Door te steunen op de combinatiefunctie $S_{\gamma^{T,F}}$ (Hoofdstuk 3) kunnen we een functie als volgt definiëren:

$$f_{S_{\gamma^{T,F}}}(\mathscr{P}_i, \mathscr{P}_j) = S_{\gamma^{T,F}}\left(\{E_O(o_1, o_2) | (o_1, o_2) \in \mathscr{P}_i \times \mathscr{P}_j\}\right). \tag{5.84}$$

De collectie van possibilistische waarheidswaarden wordt hierbij impliciet als een multiverzameling verondersteld. Deze functie kan dan worden gebruikt voor het vormen van een partitie:

$$\mathcal{H}_{f_{S_{\gamma^{T,F}}}}(O, \mathscr{B}). \tag{5.85}$$

Een dergelijke partitie is het resultaat van een Sugeno-gebaseerd herstel van transitiviteit. We kunnen nu de volgende stelling bewijzen.

Stelling 5.8
Voor een universum van objecten O en een evaluator E_O geldt er dat:

$$\forall \mathscr{P}_i \in \mathcal{H}_{f_{S_{\gamma^{T,F}}}}(O, \mathscr{B}) : \exists \mathscr{P}_j \in \mathcal{H}_{f_{\tilde{\vee}}}(O, \mathscr{B}) : \mathscr{P}_i \subseteq \mathscr{P}_j \tag{5.86}$$

Bewijs. We weten dat $\mathcal{H}_{f_{S_{\gamma^{T,F}}}}(O, \mathscr{B})$ niet altijd uniek bepaald is. Deze verschillende oplossingen ontstaan doordat de selectie van de twee klassen die samengevoegd worden niet altijd uniek bepaald is (regel 5 in het algoritme).

Echter, we kunnen inzien dat er bij elke mogelijke keuze van twee klassen $\mathscr{P}_k$ en $\mathscr{P}_l$ moet gelden:

$$\mathscr{B}\left(f_{S_{\gamma T,F}}(\mathscr{P}_k,\mathscr{P}_l)\right) \Rightarrow \mathscr{B}\left(f_{\tilde{\vee}}\ (\mathscr{P}_k,\mathscr{P}_l)\right). \tag{5.87}$$

Elke twee klassen die kunnen worden samengevoegd onder $f_{S_{\gamma T,F}}$, zullen bijgevolg zeker worden samengevoegd onder $f_{\tilde{\vee}}$. □

Stelling 5.8 heeft als gevolg dat er geldt:

$$\mathcal{H}_{f_{S_{\gamma T,F}}}(O,\mathscr{B}) = \bigcup_{\mathscr{P}_i \in \mathcal{H}_{f_{\tilde{\vee}}}\ (O,\mathscr{B})} \left(\mathcal{H}_{f_{S_{\gamma T,F}}}(\mathscr{P}_i,\mathscr{B})\right). \tag{5.88}$$

Voorbeeld 5.1
Beschouw een universum $O = \{a,b,c,d\}$ en een niet-transitieve evaluator E_O zoals getoond in Tabel 5.1. Veronderstel een beslissingsmodel $\mathscr{B}$ met $z = 0.3$. Laat ons eerst nagaan wat het resultaat is van $\mathcal{H}_{f_{\tilde{\vee}}}(O,\mathscr{B})$. Bij de berekening hiervan worden eerst b en c samen in een klasse genomen. Daarna zijn er verschillende mogelijkheden. Ofwel wordt a bij b en c gevoegd, ofwel wordt d bij b en c gevoegd, ofwel worden a en d samengebracht in een klasse. Deze keuze heeft geen invloed op het eindresultaat. Laat ons bijvoorbeeld de derde optie kiezen. We verkrijgen dan de klassen $\{a,d\}$ en $\{b,c\}$. Deze klassen worden dan samengevoegd zodat we vinden dat:

$$\mathcal{H}_{f_{\tilde{\vee}}}(O,\mathscr{B}) = \{\{a,b,c,d\}\}. \tag{5.89}$$

	a	b	c	d
a	(1,0)	(1,0.5)	(0,1)	(1,0.5)
b	(1,0.5)	(1,0)	(1,0.1)	(1,0.6)
c	(0,1)	(1,0.1)	(1,0)	(0,1)
d	(1,0.5)	(1,0.6)	(0,1)	(1,0)

Tabel 5.1: Voorbeeld van een evaluator

Laat ons nu nagaan wat het resultaat is van $\mathcal{H}_{f_{\tilde{\wedge}}}(O,\mathscr{B})$. Bij de berekening hiervan worden b en c eerst samen in een klasse geplaatst. Daarna worden a en d samen in een klasse geplaatst, waarna het algoritme stopt vermits er niet genoeg zekerheid op coreferentie is tussen c en a (of d). We vinden dus:

$$\mathcal{H}_{f_{\tilde{\wedge}}}(O,\mathscr{B}) = \{\{a,d\},\{b,c\}\}. \tag{5.90}$$

In termen van kardinaliteit bestaat er echter een grotere klasse $\{a,b,d\}$ waarvoor alle mogelijke koppels geldig zijn volgens het beslissingsmodel $\mathscr{B}$. Dit voorbeeld toont aan dat het resultaat van $\mathcal{H}_{f_{\tilde{\wedge}}}$ een partitie is die niet noodzakelijk de minimale partitie is op basis van een transitieve opening van $R_{\mathscr{B}}$ (een

	a	b	c	d
a	x	x		x
b	x	x	x	x
c		x	x	
d	x	x		x

	a	b	c	d
a	x			x
b		x	x	
c		x	x	
d	x			x

	a	b	c	d
a	x	x		x
b	x	x		x
c			x	
d	x	x		x

Figuur 5.1: $R_{\mathscr{B}}$ (links), $R'_{\mathscr{B}}$ (midden) en $R^{-}_{\mathscr{B}}$ (rechts)

maximale transitieve deelrelatie van $R_{\mathscr{B}}$). Dit wordt duidelijk door Figuur 5.1 te beschouwen, waarin van links naar rechts $R_{\mathscr{B}}$, $R'_{\mathscr{B}}$ (relatie gebruikt door $\mathcal{H}_{f\ \tilde{\wedge}}$) en $R^{-}_{\mathscr{B}}$ (transitieve opening) staan afgebeeld.

We zullen nu aantonen dat het mogelijk is om de klassen verkregen door $\mathcal{H}_{f\ \tilde{\vee}}\ (O, \mathscr{B})$ alternatief op te splitsen door iteratief een maximale en voldoende klasse af te splitsen.

Definitie 5.6 (Maximale en voldoende klasse)
Gegeven een eindige verzameling O van objecten, een evaluator E_O en een beslissingsmodel $\mathscr{B}$. Een maximale en voldoende klasse van O ten opzichte van de drempelwaarde $\eta \in [0, 1]$, is de grootste deelverzameling $[\![O]\!]$ van O zodat:

$$\left| \{(o_1, o_2) | (o_1, o_2) \in [\![O]\!] \times [\![O]\!] \wedge \mathscr{B}\left(E_O(o_1, o_2)\right) = T\} \right| \geq \eta |[\![O]\!]|^2. \qquad (5.91)$$

Wanneer we voor een eindig universum O een $|O| \times |O|$-matrix beschouwen, dan kunnen we deze matrix voorstellen als een graaf met $|O|$ knopen. Het zoeken van een maximale en voldoende klasse $[\![O]\!]$ komt in een dergelijke graaf overeen met het zoeken van een deelgraaf zodat de connectiviteit voldoende is. Dit staat ook wel bekend als het cliqueprobleem [80]. Op basis van de definitie van een maximale en voldoende klasse kunnen we de maximale en voldoende splitsing van een verzameling van objecten definiëren.

Definitie 5.7 (Maximale en voldoende splitsing)
Gegeven een eindige verzameling O van objecten, een evaluator E_O en een beslissingsmodel $\mathscr{B}$. Een maximale en voldoende splitsing van O wordt gegeven door de iteratieve functie:

$$\mathscr{H}\left(O, \mathscr{B}\right) = \begin{cases} \{O\} & \textbf{\textit{als}} \quad |O| = 1 \\ \{[\![O]\!]\} \cup \mathscr{H}\left(O \backslash [\![O]\!], \mathscr{B}\right) & \textbf{\textit{anders}}. \end{cases} \qquad (5.92)$$

Definitie 5.8 (Maximale en voldoende partitie)
Gegeven een eindige verzameling O van objecten, een evaluator E_O en een beslissingsmodel $\mathscr{B}$. Een maximale en voldoende partitie van O is gedefinieerd

als:

$$(\mathscr{H} \circ \mathcal{H}_{f_{\triangledown}})(O, \mathscr{B}) = \bigcup_{\mathscr{P}_i \in \mathcal{H}_{f_{\triangledown}}(O,\mathscr{B})} \mathscr{H}(\mathscr{P}_i). \tag{5.93}$$

Een maximaal en voldoende partitie komt tot stand door eerst een partitie te maken met $\mathcal{H}_{f_{\triangledown}}$ en daarna deze partitieklassen verder op te splitsen door maximale en voldoende splitsing. In Voorbeeld 5.1 leidt dit tot de partitie $\{\{a, b, d\}, \{c\}\}$. Merk op dat in een partitieklasse $\mathscr{P}_i$ gegenereerd door $\mathcal{H}_{f_{\triangledown}}$ steeds minstens $|\mathscr{P}_i| - 1$ koppels van verschillende objecten zitten, waarvoor de evaluatie voldoende is onder $\mathscr{B}$. Ook is elk object steeds coreferent met zichzelf, zodat er nog eens $|\mathscr{P}_i|$ koppels waarvoor de evaluatie voldoende is onder $\mathscr{B}$. Hieruit volgt dat er dus steeds $|\mathscr{P}_i| - 1 + |\mathscr{P}_i| = 2|\mathscr{P}_i| - 1$ koppels zijn waarvoor de evaluatie voldoende is. Gelet op het feit dat er in totaal $|\mathscr{P}_i|^2$ mogelijke koppels zijn, impliceert dit gegeven een ondergrens voor η:

$$\eta \in \left[\frac{2|\mathscr{P}_i| - 1}{|\mathscr{P}_i|^2}, 1\right]. \tag{5.94}$$

We merken hier op dat er geldt dat:

$$(\mathscr{H} \circ \mathcal{H}_{f_{\triangledown}})(O, \mathscr{B}) \neq \mathscr{H}(O, \mathscr{B}). \tag{5.95}$$

Het verschil met Sugeno-gebaseerd herstel is dat de zekerheid gegeven door E_O minder van belang is. Bij maximale en voldoende splitsing wordt enkel rekening gehouden met $\mathscr{B}(E_O(o_1, o_2))$, d.i. er wordt geverifieerd of de zekerheid op coreferentie voldoende is of niet. Bij Sugeno-gebaseerd herstel bepaalt de feitelijke zekerheid $1 - \mu_{E_O(o_1,o_2)}(F)$ de volgorde waarin klassen worden samengevoegd. Dit is te zien in Voorbeeld 5.1 waar de grote zekerheid op coreferentie tussen b en c er voor zorgt dat b en c in dezelfde klasse worden geplaatst. Maximale en voldoende splitsing probeert, als alternatief voor Sugeno-gebaseerd herstel, een zo groot mogelijke verzameling van objecten te vinden waarvoor onderlinge koppels met voldoende zekerheid coreferent zijn.

5.3.2 Consistentie van $R_{\mathscr{D}}^{T,F}$

Het is nu mogelijk om op basis van een evaluator E_O en een beslissingsmodel $\mathscr{B}$, een collectie van objecten O te partitioneren tot een partitie $\mathscr{P}$, waarbij elke klasse in de partitie coreferente objecten groepeert. Deze partitie komt overeen met een equivalentierelatie $R'_{\mathscr{B}}$ over O en $R'_{\mathscr{B}}$ kan worden gezien als een relatie die is afgeleid van $R_{\mathscr{B}}$ en waarbij de transitiviteit is hersteld. Bijgevolg kan $R'_{\mathscr{B}}$ worden gebruikt als een benadering van $\leftrightarrow$.

Het is reeds aangehaald dat dit ook geldt voor $R_{\mathscr{D}}^{T}$ in het geval van een driewaardig beslissingsmodel. In het geval van een driewaardig beslissingsmodel kunnen we dus eerst de relatie $R_{\mathscr{D}}^{T}$ consistent maken door de transitiviteit te herstellen. Laat ons daarom in wat volgt veronderstellen dat $R_{\mathscr{D}}^{T}$ transitief is. We weten nu dat er geldt:

$$\overline{R_{\mathscr{D}}^{T}} = R_{\mathscr{D}}^{F} \cup R_{\mathscr{D}}^{T,F}. \tag{5.96}$$

Om consistent te zijn moeten de drie relaties voldoen aan (5.40) en (5.41). Er bestaat een eenvoudige procedure om aan deze voorwaarden te voldoen. Beschouw $R^T_{\mathscr{D}}$ en de bijhorende partitie $\mathscr{P}$ van O. Voor een object o noteren we de partitieklasse waartoe o behoort als $\mathscr{P}_o$. Beschouw nu de relatie $R^{T,F}_{\mathscr{D}}$ verkregen door toepassing van het driewaardig beslissingsmodel $\mathscr{D}$ en beschouw de relatie $R'^{T,F}_{\mathscr{D}}$ zodat:

$$\forall(o_1,o_2) \in R^{T,F}_{\mathscr{D}} : \forall(o'_1,o'_2) \in \mathscr{P}_{o_1} \times \mathscr{P}_{o_2} : (o'_1,o'_2) \in R'^{T,F}_{\mathscr{D}} \tag{5.97}$$

dan geldt er onmiddellijk dat $R'^{T,F}_{\mathscr{D}}$ voldoet aan (5.40) en (5.41). Hiermee is aangetoond dat we, vertrekkende van een tweewaardig beslissingsmodel eenvoudig een driewaardig beslissingsmodel kunnen opstellen dat leidt tot consistente beslissingen. Om die reden zal in Hoofdstuk 7 enkel nog worden uitgegaan van een tweewaardig beslissingsmodel.

5.4 Inconsistenties bij gedeeltelijke coreferentie

Tot hier toe is in deze thesis voornamelijk aandacht besteed aan het coreferentieprobleem. De gedeeltelijke coreferentieproblemen (Hoofdstuk 2) zijn verwant aan het coreferentieprobleem en kunnen bijgevolg op een gelijkaardige manier worden aangepakt. Dit willen we hier nader toelichten. Eerst en vooral kunnen de definities van evaluatoren grotendeels worden behouden, met dat verschil dat de propositie voor twee objecten o_1 en o_2 wordt:

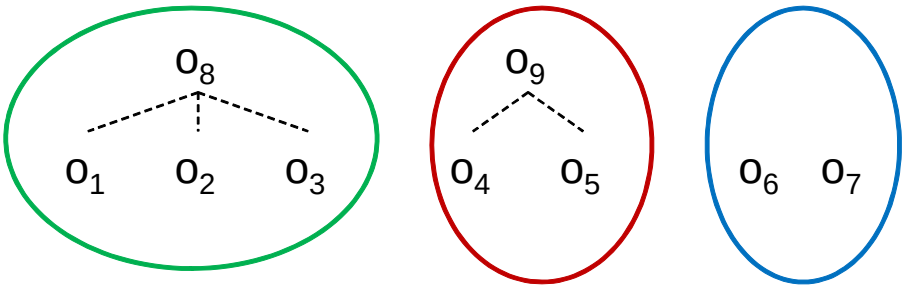
$$p_{(o_1,o_2)} = \text{“}o_1 \text{ en } o_2 \text{ zijn gedeeltelijk coreferent”}.$$

Voor syntactische evaluatoren zal de kennis die wordt gebruikt om een oordeel te vellen, typisch niet verschillen van het gewone coreferentieprobleem. Dit mag intuïtief vreemd lijken, gezien de eigenschappen reflexiviteit, symmetrie en transitiviteit van een evaluator in strijd kunnen zijn met de specifieke relatie die wordt gezocht. Zo is de relatie $\leftrightarrow_{\subseteq}$ bijvoorbeeld een partiële orderelatie over O en bijgevolg niet symmetrisch. Echter, om dezelfde reden dat transitiviteit van een evaluator niet noodzakelijk wordt afgedwongen in het geval van gewone coreferentie, wordt het niet-symmetrisch zijn van een evaluator ook niet vereist in de context van $\subseteq$-coreferentie. Symmetrie kan in dat geval als een inconsistentie worden gezien, net als het niet-transitief zijn van een evaluator bij gewone coreferentie. Dit leidt ons naar een aanpak waarbij we, ongeacht het gestelde probleem, uitgaan van een verzameling van evaluatoren, maar waarbij de relatie over O anders wordt geconstrueerd. In het geval van $\leftrightarrow_{\subseteq}$ bijvoorbeeld, kan symmetrie worden opgegeven ten voordele van transitiviteit, d.i. er kunnen koppels uit de symmetrische relatie worden verwijderd, waarna een transitieve relatie wordt verkregen. In het geval van $\leftrightarrow_{\cap}$ (d.i. een reflexieve en symmetrische relatie) is helemaal geen herstel nodig, gelet op het feit dat we steeds reflexieve en symmetrische evaluatoren beschouwen. Hoewel hierop verder niet wordt ingegaan, kan men voor gedeeltelijke coreferentie steeds een binaire relatie afleiden die voldoet aan de vereiste eigenschappen van het specifieke model,

net zoals we hier hebben gedaan voor coreferentie. Een aandachtspunt hierbij is de voorstellingswijze van de verkregen relatie. Meer bepaald, in het geval van het gewone coreferentieprobleem is de afgeleide relatie een equivalentierelatie, waardoor we een partitie over O kunnen afleiden. Een dergelijke partitie is bijzonder nuttig omwille van de eenduidige voorstellingsvorm, zodat de verdere verwerking van de resultaten eenvoudig is (Hoofdstuk 9). Dit voordeel verdwijnt wanneer de relatie geen equivalentierelatie meer is. Echter, voor de twee vermelde problemen van gedeeltelijke coreferentie ($\leftrightarrow_{\subseteq}$ en $\leftrightarrow_{\cap}$) is het steeds mogelijk een zinvolle equivalentierelatie af te leiden. De manier waarop deze relatie wordt gevormd is het onderwerp van volgende deelsecties.

5.4.1 Model voor $\leftrightarrow_{\subseteq}$-coreferentie

In het geval van $\leftrightarrow_{\subseteq}$-coreferentie kan op basis van een evaluator een relatie $R_{\mathscr{B}}$ worden afgeleid, die reflexief en symmetrisch is. Wanneer voor deze relatie een herstel van transitiviteit wordt toegepast, verkrijgen we een equivalentierelatie die modelleert dat twee objecten al dan niet tot eenzelfde deelhiërarchie behoren. Dit idee wordt voorgesteld in Figuur 5.2. In deze figuur is te zien hoe objecten o_1, o_2 en o_3 allen gedeeltelijk coreferent zijn met o_8. Deze vier objecten kunnen dan als één klasse (de groene) worden voorgesteld. Hetzelfde geldt voor andere afgebeelde objecten. Het voordeel van deze aanpak is dat voor verdere verwerking alle objecten die eenzelfde entiteit (gedeeltelijk) beschrijven bij elkaar worden geplaatst. Door het invoeren van transitiviteit gaat een deel van de structurele orde-informatie verloren, maar deze informatie kan typisch worden hersteld door technieken uit Hoofdstuk 9. We benadrukken dat de toepasbaarheid van deze aanpak sterk afhangt van de aard van het probleem. Hoe meer objecten $\subseteq$-coreferent zijn met andere objecten, hoe sterker de verkregen equivalentierelatie zal aanleunen bij $O \times O$. Er zullen dan weinig klassen zijn, zodat de resultaten weinig informatie bieden. Deze aanpak werkt wel goed voor problemen die dicht aanleunen bij het klassieke coreferentieprobleem.



Figuur 5.2: Equivalentieklassen bij $\leftrightarrow_{\subseteq}$-coreferentie

5.4.2 Model voor $\leftrightarrow_\cap$-coreferentie

Op een gelijkaardige manier kan in het geval van $\leftrightarrow_\cap$-coreferentie eveneens een equivalentierelatie worden afgeleid door de geïntroduceerde methode voor herstel van transitiviteit toe te passen. Dit leidt naar een aanpak waarbij $\leftrightarrow_\cap$-coreferentie wordt benaderd als $\leftrightarrow$-coreferentie. Een dergelijke benadering heeft zijn nut bij het vergelijken van objecten zoals beelden, video en tekst. Voor dergelijke objecten is de precieze vorm van de referentiefunctie ρ^* niet altijd makkelijk te bepalen, d.i. het is niet altijd eenduidig bepaald welke entiteiten worden beschreven. Bij het vergelijken van tekstuele documenten zullen we een aanpak schetsen in Hoofdstuk 8 die steunt op een dergelijke benadering. Opnieuw benadrukken we dat het vormen van een equivalentierelatie wel degelijk leidt tot een benaderde oplossing. Gelet op de literatuur over een vage aanpak van clusteren is het duidelijk dat deze benadering niet altijd wenselijk is. Immers, de hele idee achter een vage aanpak voor clusteren is dat een object tot verschillende clusters kan behoren, waardoor een afhankelijkheidsrelatie een correcter beeld geeft van de gezochte clusters dan een equivalentierelatie.

5.5 Kardinaliteitsrestricties

Daar waar in de vorige sectie is aangetoond hoe een verzameling van objecten kan worden gepartitioneerd, onderzoeken we in deze sectie hoe randvoorwaarden met betrekking tot de kardinaliteit van partitieklassen kunnen worden afgedwongen. Meer bepaald zijn we in deze sectie geïnteresseerd in het zogenaamde identificatieprobleem. Bij het identificatieprobleem beschouwt men twee groepen van objecten, $G_1 \subset O$ en $G_2 \subset O$, waarbij een object uit G_1 met ten hoogste één object uit G_2 coreferent kan zijn en omgekeerd. Een voor de hand liggend voorbeeld van het identificatieprobleem is de identificatie van personen op basis van vingerafdrukken, DNA of de oorbiometrie, maar ook bij het vergelijken van twee onafhankelijke bronnen gaat de veronderstelling van het identificatieprobleem vaak op. Zo is de achterliggende databank van websites vaak vrij van coreferente objecten, zodat bij de vergelijking van twee onafhankelijke websites de één-op-één vereiste verondersteld mag worden. Oplossen van het coreferentieprobleem betekent in dit geval het construeren van een één-op-één afbeelding tussen G_1 en G_2. De partitie $\mathscr{P}$ uit vorige sectie moet dan bestaan uit partitieklassen met ten hoogste twee objecten.

Stel een evaluator E_O en stel drie objecten $o_1 \in G_1$, $o_2 \in G_2$ en $o_3 \in G_2$. Beschouw een beslissingsmodel $\mathscr{B}$ waarvoor geldt:

$$\mathscr{B}\left(E_O(o_1, o_2)\right) \wedge \mathscr{B}\left(E_O(o_1, o_3)\right) \tag{5.98}$$

dan kunnen o_1, o_2 en o_3 tot dezelfde klasse behoren na uitvoering van $\mathcal{H}_f$. In dat geval moet volgens de één-op-één vereiste ofwel o_2 ofwel o_3 uit de klasse worden verwijderd, zodat een keuze zich opdringt. Voor deze keuze baseren we ons op de zekerheid die E_O biedt over de coreferentie van enerzijds o_1 met o_2 en anderzijds o_1 met o_3, waarbij het koppel met de meeste zekerheid

behouden blijft. Uiteraard is het mogelijk dat $E_O(o_1, o_2) = E_O(o_1, o_3)$, zodat we in dat geval moeten kijken of er nog andere objecten tot de klasse behoren. Als bijvoorbeeld ook geldt dat $\mathscr{B}(E_O(o_3, o_4))$ maar niet $\mathscr{B}(E_O(o_2, o_4))$, dan kunnen we o_1 afbeelden op o_2 en o_3 op o_4. Dit betekent dat we een leximax-optimale afbeelding tussen G_1 en G_2 moeten construeren. De constructie van een dergelijke afbeelding is uitgelegd in Hoofdstuk 4 in de context van het vergelijken van collecties. Een belangrijke vaststelling is dat de aanpak uit Hoofdstuk 4 geen garantie biedt op een unieke oplossing. Dit is inherent aan het probleem en hiermee moet rekening worden gehouden bij de beoordeling van resultaten. Experimenten gerapporteerd in Hoofdstuk 6 tonen aan dat het opleggen van beperkingen op de kardinaliteit van partitieklassen leidt tot een significante verbetering in de oplossing van een coreferentieprobleem.

5.6 Conclusie

In dit hoofdstuk is onderzocht hoe een consistente oplossing voor het coreferentieprobleem kan worden gevonden op basis van een evaluator. Hiervoor zijn twee soorten beslissingsmodellen geïntroduceerd: tweewaardige en driewaardige beslissingsmodellen. Dit zijn functies die possibilistische waarheidswaarden afbeelden op respectievelijk de Boolese ruimte en de machtsverzameling van de Boolese ruimte. Elk tweewaardig beslissingsmodel $\mathscr{B}$ geeft aanleiding tot een binaire relatie $R_{\mathscr{B}}$ die koppels van coreferente objecten bevat. Willen we $R_{\mathscr{B}}$ gebruiken als benadering van de coreferentierelatie $\leftrightarrow$, dan moet $R_{\mathscr{B}}$ reflexief, symmetrisch en transitief zijn. Het is aangetoond dat transitiviteit van $R_{\mathscr{B}}$ enkel voorkomt uit een transitieve evaluator. Vermits evaluatoren meestal niet transitief zijn, is onderzocht hoe uit $R_{\mathscr{B}}$ een transitieve relatie kan worden afgeleid die de kennis van de gebruikte evaluator zo goed mogelijk respecteert. Dit leidt tot een Sugeno-gebaseerd herstel van transitiviteit. Naast deze Sugeno-gebaseerde aanpak is een alternatieve aanpak op basis van maximale en voldoende klassen voorgesteld. De ingevoerde methode kan tevens worden toegepast in het geval van een driewaardig beslissingsmodel op de relatie $R^T_{\mathscr{D}}$. Bijkomend moet voor een driewaardig beslissingsmodel de relatie $R^{T,F}_{\mathscr{D}}$ aan consistentievoorwaarden voldoen. Het is aangetoond hoe deze voorwaarden kunnen worden afgedwongen. Eens consistente relaties verkregen zijn, kan het universum van objecten worden gepartitioneerd in klassen van coreferente objecten. Er is bijkomend aangetoond hoe deze partitieklassen geïnterpreteerd moeten worden in een context van gedeeltelijke coreferentie. Ten slotte is onderzocht hoe de randvoorwaarden van het identificatieprobleem kunnen worden toegepast op de uitkomst van een evaluator. Hiervoor wordt gesteund op de constructie van een leximax-optimale afbeelding (Hoofdstuk 4).

Hoofdstuk 6

Evaluatoren voor strings

6.1 Inleiding

In Hoofdstuk 3 is onderzocht hoe possibilistische waarheidswaarden gecombineerd kunnen worden wanneer verschillende actoren kennis beschrijven over eenzelfde Boolese propositie p. De resulterende possibilistische waarheidswaarde geeft dan de kennis die voorhanden is over p. Deze methode is toegepast in Hoofdstuk 4 voor de constructie van evaluatoren voor collecties. De resultaten uit Hoofdstuk 4 zullen op hun beurt in dit hoofdstuk worden toegepast voor de constructie van evaluatoren voor karakterstrings, ook kortweg strings genoemd. Na deze inleiding zullen we in Sectie 6.2 enkele bestaande technieken voor stringvergelijking bespreken. Vervolgens worden enkele basisconcepten geïntroduceerd in Sectie 6.3. Deze basisconcepten worden gebruikt om een eerste type evaluator voor strings te definiëren in Sectie 6.4. In Sectie 6.5 wordt een tweede type gedefinieerd door strings te transformeren naar een collectie van deelstrings. Er worden enkele geavanceerde mogelijkheden voor evaluatie van strings onderzocht in Sectie 6.6. Daarna wordt in Sectie 6.7 bestudeerd hoe de relevante parameters van een evaluator bepaald kunnen worden. Sectie 6.8 biedt een uitgebreide bespreking van uitgevoerde experimenten. Sectie 6.9 geeft een overzicht van de belangrijkste bevindingen.

In dit hoofdstuk zullen de letters s, t en r worden gebruikt om willekeurige strings voor te stellen. Hoewel in Hoofdstuk 1 de notaties t en s gebruikt zijn als notatie voor respectievelijk een willekeurige triangulaire norm en een willekeurige triangulaire conorm, zal uit de context steeds blijken of we met s en t een variabele, dan wel een functie bedoelen.

Om deze inleiding te besluiten, willen we kort aandacht besteden aan de interpretatie die aan strings moet worden gehecht in het kader van dit hoofdstuk. Een string wordt gebruikt om tekstuele data voor te stellen waarbij, zonder verdere beperkingen, de inhoud relatief uiteenlopend kan zijn. In dit hoofdstuk wordt een evaluator voor strings gedefinieerd vanuit het standpunt van complexe objecten die entiteiten beschrijven door een goed gestructureerde

beschrijving van afgelijnde eigenschappen van de entiteit. Er wordt bijgevolg verondersteld dat een string wordt gebruikt om een antwoord te formuleren op de vraag: "Wat is eigenschap a van entiteit e?", waardoor de inhoud van een string typisch een relatief kort en duidelijk antwoord is dat niet bestaat uit volzinnen. Hierdoor moet men geen rekening houden met grammaticale regels van taal. Andere situaties zullen verderop in deze thesis worden onderzocht. Zo zal in Hoofdstuk 7 worden aangetoond dat een complex object kan worden herleid tot een string, waardoor vergelijking van complexe objecten kan worden herleid tot vergelijking van strings. In dit geval vormt een string een antwoord op de vraag: "Wat zijn eigenschappen a_1,..., a_n van entiteit e?". In Hoofdstuk 8 wordt nog een ander probleem bestudeerd waarbij een string een tekstueel document voorstelt. In een dergelijke situatie wordt een entiteit niet langer beschreven als een opsomming van afgelijnde eigenschappen en is een string het resultaat van de vraag: "Hoe zou u entiteit e beschrijven". In een dergelijke context bestaat een string wel degelijk uit een opeenvolging van volzinnen, die een taalafhankelijke grammatica respecteren.

6.2 Overzicht van de literatuur

In de literatuur over stringvergelijking wordt onderscheid gemaakt tussen drie klassen van methoden voor het vergelijken van strings: karaktergebaseerde methoden, woordgebaseerde methoden en fonetische methoden. In dit overzicht willen we de belangrijkste resultaten van elk van deze drie klassen kort bespreken.

Een eerste klasse van methoden is deze van de karaktergebaseerde methoden. De oudste methode in deze klasse is de editeerafstand van Levenshtein [81], ook wel de Levenshtein afstand genoemd. De Levenshtein afstand tussen twee strings s en t is het minimaal aantal editeeroperaties, dat nodig is om s te transformeren naar t. De mogelijke operaties hierbij zijn het invoegen van een karakter, het verwijderen van een karakter en het vervangen van een karakter door een ander. Belangrijk hierbij is dat de Levenshtein afstand voldoet aan de voorwaarden van een metriek (niet-negatief, reflexief, symmetrisch en driehoeksongelijkheid) zodat de benaming afstand gerechtvaardigd is. Damerau stelt in [82] een gelijkaardige editeerafstand voor, maar beschouwt één extra editeeroperatie, namelijk transpositie (d.i. het verwisselen van twee naburige karakters). Damerau stelt in zijn werk dat zijn verzameling van operaties overeenkomt met meer dan 80% van alle menselijke schrijffouten. Enkele jaren na de invoering van de editeerafstand wordt het vergelijken van strings een relevant onderwerp in biomedische kringen wanneer Needleman en Wunsch [83] een similariteitsmaat definiëren voor proteïnesequenties. De basis van hun aanpak is het zoeken naar de *langste gemeenschappelijke deelsequentie*. Smith en Waterman verfijnen deze methode door rekening te houden met lokale alignaties [84]. Monge en Elkan zien in dat de voorgestelde algoritmen voor proteïnesequenties bijzonder nuttig zijn in de context van natuurlijke talen [85]. Zij passen de editeerafstand aan door editeeroperaties toepasbaar te maken op blokken

van karakters, eerder dan uitsluitend karakters. Monge en Elkan stellen dat een dergelijk model beter rekening houdt met afkortingen. Een interessante techniek is deze van Jaro [86]. Dit is een heuristische methode die bijzonder geschikt is voor korte strings en die een significant lagere complexiteit heeft dan de reeds vermelde technieken. Winkler heeft later de aanpak van Jaro uitgebreid waarbij hij in rekening brengt dat de waarschijnlijkheid op een fout aan het begin van een string lager is dan aan het einde van een string. In de aanpak van Winkler worden verschillen tussen strings vooraan bijgevolg sterker afgestraft dan verschillen achteraan [87]. Een vergelijkende studie[1] stelt dat de aanpak van Monge en Elkan de beste gemiddelde accuraatheid vertoont, dicht gevolgd door de aanpak van Jaro [88]. Dezelfde studie toont echter ook aan dat de aanpak van Monge en Elkan een grootteorde trager is dan de aanpak van Jaro. Ook wordt ondervonden dat de aanpak van Monge en Elkan eerder probleemspecifiek is en niet robuust. De studie vermeldt dat de aanpak van Jaro een hoge accuraatheid combineert met een lage complexiteit, waardoor het een bijzonder interessante aanpak is.

Een tweede klasse van methoden wordt gevormd door de woordgebaseerde methoden. Bij elk van deze methoden wordt een string opgesplitst in woorden door een splitsingsfunctie of een blokfunctie[2]. Bij de meeste methoden worden de verkregen woorden voorgesteld in een vectorruimte en wordt elke string afgebeeld op een vector van deelstrings, waardoor het probleem van stringvergelijking wordt vertaald naar een probleem van vectorvergelijking. Vergelijking van strings gebeurt dan door vergelijking van de bijhorende vectoren, bijvoorbeeld door de cosinus van de hoek tussen beide vectoren te berekenen [89]:

$$\frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \, \|\mathbf{y}\|}. \tag{6.1}$$

Meestal wordt een mechanisme gebruikt dat aan de verschillende vectordimensies gewichten toekent. Het meest gekende voorbeeld hiervan is het TFIDF schema [89]. In dit schema wordt voor een string s het gewicht van een deelstring t berekend als het product van (1) de frequentie van t binnen s en (2) de logaritme van de omgekeerde frequentie van t binnen de datacollectie. Daar waar karaktergebaseerde methoden geschikt zijn voor het omgaan met fouten op karakterniveau, zijn woordgebaseerde methoden geschikt voor het omgaan met fouten op woordniveau. De studie in [88] toont aan dat woordgebaseerde methoden een grootteorde sneller zijn dan de snelste karaktergebaseerde methoden. Om het beste van twee werelden te combineren is recent onderzoek gedaan naar twee-niveau systemen [85, 90, 91]. Dit is een woordgebaseerde methode (hoogste niveau), waarbij woorden vergeleken worden met behulp van een karaktergebaseerde methode (laagste niveau). Dergelijke systemen blijken bijzonder accuraat te zijn [88]. De techniek SoftTFIDF, waarbij de Jaro-Winkler methode wordt gebruikt op het laagste niveau, komt naar voren als de beste

[1] Deze studie maakt gebruik van de implementatie die te vinden is op de website http://secondstring.sourceforge.net

[2] In de meest algemene context worden taalkundige woordgrenzen niet in rekening gebracht en beschouwt men blokken van karakters.

methode met betrekking tot gemiddelde accuraatheid. De complexiteit van een dergelijke methode hangt af van enerzijds het vergelijken van vectoren en anderzijds de complexiteit van de techniek gebruikt op het laagste niveau. In dit hoofdstuk zal een dergelijk twee-niveau systeem worden onderzocht in een possibilistische context door te steunen op de evaluator voor collecties die in Hoofdstuk 4 is geïntroduceerd.

De laatste klasse betreft de fonetische methoden. Daar waar de eerste twee stromingen oplossingen bieden voor fouten te wijten aan een niet-homogene schrijfwijze, proberen fonetische methoden rekening te houden met de uitspraak van woorden. De eerste techniek in deze stroming wordt toegeschreven aan Russell en dateert uit 1918 [92]. De gepatenteerde aanpak van Russell bestaat erin elk karakter te transformeren naar ofwel een numerieke code, ofwel een scheidingssymbool. Russell legt dan een aantal stappen vast om te komen tot een globale numerieke code, zodat gelijkaardig uitgesproken strings worden afgebeeld op gelijke codes. Een studie van Newcombe [93] rapporteert dat de Soundex-code van Russell constant blijft bij ongeveer 66% van de geobserveerde spellingsvariaties in medische databanken. Newcombe besluit daaruit dat de toepasbaarheid van Russell's aanpak niet beperkt blijft tot enkel fonetische variaties. Een groot voordeel van de Soundex-code is dat de transformatie van strings naar codes een lineaire complexiteit heeft. Het vergelijken van codes kan bijzonder efficiënt gebeuren door gebruik te maken van indices en sorteeralgoritmen. We willen terzijde opmerken dat de Soundex-code de enige techniek is, die standaard is opgenomen in de SQL-dialecten van enerzijds grote databanksystemen als Oracle en SQLServer en anderzijds *Open Bron* systemen als MySql. Naast de Soundex-code bestaan andere technieken, die enkel verschillen van Russell's aanpak in de manier waarop codering van strings gebeurt [94, 95, 96]. In het recentere werk van Philips [97] wordt voor het eerst een aanpak voorgesteld waarbij meerdere coderingen mogelijk zijn.

6.3 Definities

In deze sectie worden een aantal definities en basisoperatoren omtrent strings geïntroduceerd. We veronderstellen het bestaan van een eindig alfabet $\mathscr{A}$. De elementen van het alfabet worden karakters genoemd. Een string kan worden gedefinieerd als een sequentie van karakters.

Definitie 6.1 (String)
Een string s met lengte n over een alfabet $\mathscr{A}$ wordt gekarakteriseerd door de functie τ_s:

$$\tau_s : I_n \rightarrow \mathscr{A} \tag{6.2}$$

waarbij $I_n = \{1, ..., n\}$ de indexverzameling wordt genoemd. Hierbij is $\tau_s(i)$ gelijk aan het i^{de} karakter van de string s. Een string met lengte 0 wordt een lege string genoemd en wordt genoteerd met het symbool σ. De verzameling van alle strings wordt genoteerd als $\mathcal{S}$ en de verzameling van alle strings met lengte n wordt genoteerd als $\mathcal{S}_n$.

De lengte van een string s wordt genoteerd als $|s|$ en is bij definitie gelijk aan $|I_n|$. Twee strings s en t zijn gelijk als zowel het domein als het beeld van de functies τ_s en τ_t gelijk zijn:

$$s = t \Leftrightarrow \left(\left(I_{|s|} = I_{|t|}\right) \wedge \left(\forall i \in I_{|s|} : \tau_s(i) = \tau_t(i)\right)\right). \tag{6.3}$$

Definitie 6.2 (Indexfunctie)
De indexfunctie is gedefinieerd als een functie:

$$\mathrm{ind}_k : \mathcal{S} \times \mathscr{A} \to \mathbb{N} : (s, a) \mapsto \min\left\{i | i \in \{k+1, ..., |s|\} \wedge \tau_s(i) = a\right\} \tag{6.4}$$

waarbij er moet gelden dat $k \in \{0, ..., |s|\}$. Als karakter a niet voorkomt in s op een indexpositie groter dan k, is $\mathrm{ind}_k(s, a)$ bij conventie gelijk aan 0:

$$(\forall i \in \{k+1, ..., |s|\} : \tau_s(i) \neq a) \Rightarrow (\mathrm{ind}_k(s, a) = 0). \tag{6.5}$$

De functie ind zoekt in een string s naar de kleinste index van een bepaald karakter a die groter is dan k. Wanneer $k = 0$ betekent dit dat ind de kleinste index van een bepaald karakter a geeft. Er geldt bijgevolg dat:

$$\forall s \in \mathcal{S} : \forall a \in \mathscr{A} : \mathrm{ind}_0(s, a) = \min\left\{i | i \in I_{|s|} \wedge \tau_s(i) = a\right\}. \tag{6.6}$$

Om die reden noteren we $\mathrm{ind}_0(s, a) = \mathrm{ind}(s, a)$. Vermits een string kan worden gezien als een opeenvolging van karakters, is een interessante operatie het selecteren en herschikken van karakters, met een nieuwe string als resultaat. Een dergelijke selectie en herschikking van karakters wordt hier een transformatie van een string genoemd. De transformatie van een string s wordt gestuurd door een transformatievector $\mathbf{q}$. Deze vector bevat indexen zodat het i^{de} karakter van de nieuwe string gelijk wordt aan het $(\mathbf{q}_i)^{de}$ karakter van s. De transformatieoperator voor strings [.] wordt gedefinieerd als volgt.

Definitie 6.3 (Transformatieoperator voor strings)
Gegeven een string $s \in \mathcal{S}_n$ met indexverzameling I_n. Veronderstel een vector $\mathbf{q} = (\mathbf{q}_1, ..., \mathbf{q}_k)$ met $k \leq n$ en zodat voor alle i en j, $\mathbf{q}_i \in I_n \wedge \mathbf{q}_j \in I_n \wedge i \neq j \Rightarrow \mathbf{q}_i \neq \mathbf{q}_j$. De transformatie van s onder de vector $\mathbf{q}$ is een string $s[\mathbf{q}] \in \mathcal{S}_k$ zodat:

$$\forall i \in \{1, ..., k\} : \tau_{s[\mathbf{q}]}(i) = \tau_s(\mathbf{q}_i). \tag{6.7}$$

De transformatieoperator [.] leidt voor welbepaalde keuzes van de transformatievector $\mathbf{q}$ tot bijzondere gevallen. Voor een string s met $|\mathbf{q}| = |s|$ en $\mathbf{q}_i = i$ geldt er dat:

$$s[\mathbf{q}] = s. \tag{6.8}$$

Als er geldt dat $|\mathbf{q}| = |s|$ dan wordt $s[\mathbf{q}]$ een permutatie van s genoemd. Als er voor alle $i \in \{1, ..., k-1\}$ geldt dat $\mathbf{q}_{i+1} - \mathbf{q}_i = 1$ dan wordt de getransformeerde string een deelstring van s genoemd.

Definitie 6.4 (Deelstring)
Gegeven twee strings $s \in \mathcal{S}$ en $t \in \mathcal{S}$, dan is t een deelstring van s, genoteerd als $t \sqsubset s$, als er een transformatievector $\mathbf{q}$ van lengte $|t|$ bestaat waarvoor er geldt:

$$\forall i \in \{1, ..., |t| - 1\} : \mathbf{q}_{i+1} - \mathbf{q}_i = 1 \tag{6.9}$$

zodat $t = s[\mathbf{q}]$.

Wanneer de voorwaarde op de elementen van de transformatievector wordt verzwakt van stijgend met verschil 1 naar enkel stijgend, impliceert dit een verzwakking van het concept 'deelstring'. Het resultaat wordt dan ook een zwakke deelstring genoemd.

Definitie 6.5 (Zwakke deelstring)
Gegeven twee strings $s \in \mathcal{S}$ en $t \in \mathcal{S}$, dan is t een zwakke deelstring van s, genoteerd als $t \mathrel{\widehat{\sqsubset}} s$ als er een transformatievector $\mathbf{q}$ van lengte $|t|$ bestaat waarvoor er geldt:

$$\forall i \in \{1, ..., |t| - 1\} : \mathbf{q}_{i+1} > \mathbf{q}_i \tag{6.10}$$

zodat $t = s[\mathbf{q}]$.

Vanuit een formeel standpunt zijn (zwakke) deelstrings van s niets anders dan selecties op de indexverzameling van s. In de literatuur gebruikt met vaak de term 'deelsequentie' om aan te duiden wat hier een zwakke deelstring wordt genoemd. Er wordt bewust voor de naam 'zwakke deelstring' gekozen gezien het feit dat $s \sqsubset t \Rightarrow s \mathrel{\widehat{\sqsubset}} t$ en gezien hun sterk gelijkaardige afleiding via de transformatieoperator.

Voorbeeld 6.1
Beschouw de strings s="justine", t="tine" en r="jusine". Dan geldt er dat:

$$t \sqsubset s \tag{6.11}$$
$$t \mathrel{\widehat{\sqsubset}} s \tag{6.12}$$
$$r \mathrel{\widehat{\sqsubset}} s. \tag{6.13}$$

Een volgende interessante operatie is het concateneren van strings.

Definitie 6.6 (Concatenatie van strings)
De concatenatie van twee strings $s \in \mathcal{S}$ en $t \in \mathcal{S}$ wordt genoteerd als $s \oplus t$ en is een string $r \in \mathcal{S}_{|s|+|t|}$ zodat τ_r voldoet aan:

$$\forall i \in I_{|s|+|t|} : \tau_r(i) = \begin{cases} \tau_s(i), & i \leq |s| \\ \tau_t(i - |s|), & i > |s|. \end{cases} \tag{6.14}$$

Voorbeeld 6.2
Beschouw de strings s="justine", t=" henin". De concatenatie van deze strings is $s \oplus t$ ="justine henin".

Concatenatie van strings bezit de volgende eigenschappen:

$$(s \oplus t) \oplus r = s \oplus (t \oplus r) \tag{6.15}$$
$$s \oplus \sigma = s \tag{6.16}$$
$$\sigma \oplus s = s. \tag{6.17}$$

De componenten waaruit het resultaat van een stringconcatenatie is opgebouwd, kunnen worden gevonden met behulp van de transformatieoperator. Meer bepaald, als $r = s \oplus t$, dan geldt er:

$$s = r[\mathbf{q}^s] = s[\mathbf{q}^s] \tag{6.18}$$

waarbij $|\mathbf{q}^s| = |s|$ en $\mathbf{q}^s_i = i$. Er geldt ook dat:

$$t = r[\mathbf{q}^t] \tag{6.19}$$

waarbij $|\mathbf{q}^t| = |t|$ en $\mathbf{q}^t_i = |s| + i$. Voor twee verzamelingen van strings S en T noteren we:

$$S \oplus T = \{s \oplus t | (s,t) \in S \times T\}. \tag{6.20}$$

Eigenschap 6.1
Voor drie willekeurige strings s, t, r geldt er dat:

$$\left(s \mathrel{\widehat{\sqsubset}} t\right) \Rightarrow \left((s \oplus r) \mathrel{\widehat{\sqsubset}} (t \oplus r)\right) \tag{6.21}$$
$$\left(s \mathrel{\widehat{\sqsubset}} t\right) \Rightarrow \left((r \oplus s) \mathrel{\widehat{\sqsubset}} (r \oplus t)\right). \tag{6.22}$$

Bewijs. Als $s \mathrel{\widehat{\sqsubset}} t$, dan bestaat er een transformatievector $\mathbf{q}$ waarvoor er geldt dat $\mathbf{q}_i \leq \mathbf{q}_{i+1}$ zodat $s = t[\mathbf{q}]$. Beschouw nu de vector $\mathbf{q}^{left}$ met lengte $|r| + |s|$ waarbij:

$$\forall i \in \{1, ..., |r| + |s|\} : \mathbf{q}^{left}_i = \begin{cases} i & \textbf{als} \quad i \leq |r| \\ \mathbf{q}_i + |r| & \textbf{als} \quad i > |r|. \end{cases} \tag{6.23}$$

Er geldt dan dat $\mathbf{q}^{left}_i \leq \mathbf{q}^{left}_{i+1}$ en bovendien geldt er dat:

$$(r \oplus s) = (r \oplus t)[\mathbf{q}^{left}] \tag{6.24}$$

waaruit volgt dat:

$$(r \oplus s) \mathrel{\widehat{\sqsubset}} (r \oplus t). \tag{6.25}$$

Beschouw nu de vector $\mathbf{q}^{right}$ van lengte $|s| + |r|$ waarbij:

$$\forall i \in \{1, ..., |s| + |r|\} : \mathbf{q}^{right}_i = \begin{cases} \mathbf{q}_i & \textbf{als} \quad i \leq |s| \\ |s| + i & \textbf{als} \quad i > |s|. \end{cases} \tag{6.26}$$

Er geldt dan dat $\mathbf{q}^{right}_i \leq \mathbf{q}^{right}_{i+1}$ en bovendien geldt er dat:

$$(s \oplus r) = (t \oplus r)[\mathbf{q}^{right}] \tag{6.27}$$

waaruit volgt dat:

$$(s \oplus r) \;\widehat{\sqsubset}\; (t \oplus r)\,. \tag{6.28}$$

□

Merk op dat Eigenschap 6.1 niet noodzakelijk geldt met betrekking tot $\sqsubset$.

Een belangrijke operatie voor strings is de doorsnede. De doorsnede van twee strings s en t is de verzameling van alle deelstrings van s en t met een maximale lengte. We noemen dit de doorsnede van strings omdat de doorsnede van twee verzamelingen eveneens gelijk is aan de grootste deelverzameling in termen van kardinaliteit. Eenzelfde reden wordt aangehaald door Zadeh bij de definitie van de doorsnede van twee vaagverzamelingen [6]. Het enige verschil tussen strings en verzamelingen is dat de doorsnede van strings niet noodzakelijk uniek bepaald is. Het resultaat is hierdoor een verzameling van strings in plaats van een string.

Definitie 6.7 (Doorsnede van strings)
Voor twee strings $s \in \mathcal{S}$ en $t \in \mathcal{S}$ is de doorsnede, genoteerd als $s \sqcap t$, gelijk aan de verzameling van alle strings met lengte m waarvoor:

$$\forall r \in (s \sqcap t) \quad : \quad r \sqsubset s \wedge r \sqsubset t \tag{6.29}$$

$$\forall u \in \mathcal{S} \backslash (s \sqcap t) \quad : \quad (|u| \geq m) \Rightarrow \neg(u \sqsubset s \wedge u \sqsubset t). \tag{6.30}$$

Het getal m wordt de karakteristieke lengte van de doorsnede genoemd en wordt genoteerd als $\text{len}(s \sqcap t)$.

Voorbeeld 6.3
Beschouw de strings s=“justine”, t=“tine” en r=“jusine”. Dan geldt er dat:

$$s \sqcap t = \{\text{“tine”}\} \tag{6.31}$$

$$s \sqcap r = \{\text{“jus”}, \text{“ine”}\} \tag{6.32}$$

$$t \sqcap r = \{\text{“ine”}\}. \tag{6.33}$$

Merk op dat $s \sqcap t$ altijd minstens één string bevat. Als s en t geen enkel karakter gemeenschappelijk hebben bevat $s \sqcap t$ de lege string σ. De doorsnede van strings voldoet aan de volgende eigenschappen:

$$\text{(Commutativiteit)} \qquad s \sqcap t = t \sqcap s \tag{6.34}$$

$$\text{(Idempotentie)} \qquad s \sqcap s = \{s\} \tag{6.35}$$

$$\text{(Randvoorwaarde)} \quad \text{len}(s \sqcap t) \leq \min(|s|, |t|)\,. \tag{6.36}$$

Een volgende interessante operatie is het verschil van twee strings.

Definitie 6.8 (Verschil van strings)
Voor twee strings $s \in \mathcal{S}$ en $t \in \mathcal{S}$ is het verschil, genoteerd als $s \ominus t$, een verzameling van strings van lengte $m = |s| - \text{len}(s \sqcap t)$ zodat:

$$\forall r \in (s \ominus t) : \exists u \in (s \sqcap t) : u = s[\mathbf{q}] \wedge r = s[\mathbf{q}'] \tag{6.37}$$

waarbij $\mathbf{q}'$ *een transformatievector is die het complement is van* $\mathbf{q}$ *met betrekking tot* $(1, 2, ..., |s|)$ *en* $\mathbf{q}'_i < \mathbf{q}'_{i+1}$. *Het getal* m *wordt de karakteristieke lengte van het verschil genoemd en wordt genoteerd als* $\text{len}(s \ominus t)$.

Voorbeeld 6.4
Beschouw de strings s="justine", t="tine" en r="jusine". Dan geldt er dat:

$$s \ominus t = \{\text{“jus”}\} \tag{6.38}$$
$$s \ominus r = \{\text{“tine”, “just”}\} \tag{6.39}$$
$$t \ominus s = \{\sigma\} \tag{6.40}$$
$$t \ominus r = \{\text{“t”}\} \tag{6.41}$$
$$r \ominus s = \{\text{“ine”, “jus”}\} \tag{6.42}$$
$$r \ominus t = \{\text{“jus”}\}. \tag{6.43}$$

Het verschil van strings is geen commutatieve operator. Merk op dat voor twee strings s en t geldt dat:

$$|s \sqcap t| = |s \ominus t| = |t \ominus s|. \tag{6.44}$$

Dit reflecteert dat voor elke mogelijke string in de doorsnede een string in het verschil bestaat. De operatoren doorsnede en verschil kunnen worden verzwakt door het gebruik van $\sqsubset$ te vervangen door $\widehat{\sqsubset}$.

Definitie 6.9 (Zwakke doorsnede van strings)
Voor twee strings $s \in \mathcal{S}$ *en* $t \in \mathcal{S}$ *is de zwakke doorsnede, genoteerd als* $s \mathbin{\widehat{\sqcap}} t$, *gelijk aan de verzameling van alle strings met lengte* m *waarvoor:*

$$\forall r \in \left(s \mathbin{\widehat{\sqcap}} t\right) \quad : \quad r \mathbin{\widehat{\sqsubset}} s \wedge r \mathbin{\widehat{\sqsubset}} t \tag{6.45}$$
$$\forall u \in \mathcal{S} \backslash \left(s \mathbin{\widehat{\sqcap}} t\right) \quad : \quad (|u| \geq m) \Rightarrow \neg(u \mathbin{\widehat{\sqsubset}} s \wedge u \mathbin{\widehat{\sqsubset}} t). \tag{6.46}$$

Het getal m *wordt de karakteristieke lengte van de zwakke doorsnede genoemd en wordt genoteerd als* $\text{len}(s \mathbin{\widehat{\sqcap}} t)$.

Voorbeeld 6.5
Beschouw de strings s="justine", t="tine" en r="jusine". Dan geldt er dat:

$$s \mathbin{\widehat{\sqcap}} t = \{\text{“tine”}\} \tag{6.47}$$
$$s \mathbin{\widehat{\sqcap}} r = \{\text{“jusine”}\} \tag{6.48}$$
$$t \mathbin{\widehat{\sqcap}} r = \{\text{“ine”}\}. \tag{6.49}$$

Definitie 6.10 (Zwak verschil van strings)
Voor twee strings $s \in \mathcal{S}$ *en* $t \in \mathcal{S}$ *is het zwak verschil* $s \mathbin{\widehat{\ominus}} t$ *een verzameling van strings van lengte* $m = |s| - \text{len}(s \mathbin{\widehat{\sqcap}} t)$ *zodat:*

$$\forall r \in (s \mathbin{\widehat{\ominus}} t) : \exists u \in (s \mathbin{\widehat{\sqcap}} t) : u = s[\mathbf{q}] \wedge r = s[\mathbf{q}'] \tag{6.50}$$

waarbij $\mathbf{q}'$ *een transformatievector is die het complement is van* $\mathbf{q}$ *met betrekking tot* $(1, 2, ..., |s|)$ *en* $\mathbf{q}'_i < \mathbf{q}'_{i+1}$. *Het getal* m *wordt de karakteristieke lengte van het zwak verschil genoemd en wordt genoteerd als* $\text{len}(s \mathbin{\widehat{\ominus}} t)$.

Voorbeeld 6.6
Beschouw de strings s="justine", t="tine" en r="jusine". Dan geldt er dat:

$$
\begin{aligned}
s \mathbin{\widehat{\ominus}} t &= \{\text{“}jus\text{”}\} && (6.51)\\
s \mathbin{\widehat{\ominus}} r &= \{\text{“}t\text{”}\} && (6.52)\\
t \mathbin{\widehat{\ominus}} s &= \{\sigma\} && (6.53)\\
t \mathbin{\widehat{\ominus}} r &= \{\text{“}t\text{”}\} && (6.54)\\
r \mathbin{\widehat{\ominus}} s &= \{\sigma\} && (6.55)\\
r \mathbin{\widehat{\ominus}} t &= \{\text{“}jus\text{”}\}. && (6.56)
\end{aligned}
$$

Alle eigenschappen voor $\sqcap$ die hierboven vermeld zijn, gelden ook voor $\mathbin{\widehat{\sqcap}}$. Bijkomend geldt er distributiviteit van $\oplus$ over $\mathbin{\widehat{\sqcap}}$.

Eigenschap 6.2 (Distributiviteit van $\oplus$ over $\mathbin{\widehat{\sqcap}}$)
Voor drie strings s_1, s_2 en t geldt er dat:

$$
\begin{aligned}
\{t\} \oplus \left(s_1 \mathbin{\widehat{\sqcap}} s_2\right) &= (t \oplus s_1) \mathbin{\widehat{\sqcap}} (t \oplus s_2) && (6.57)\\
\left(s_1 \mathbin{\widehat{\sqcap}} s_2\right) \oplus \{t\} &= (s_1 \oplus t) \mathbin{\widehat{\sqcap}} (s_2 \oplus t) && (6.58)
\end{aligned}
$$

Bewijs. Gelet op Eigenschap 6.1 geldt er voor $r \in \left(s_1 \mathbin{\widehat{\sqcap}} s_2\right)$ dat:

$$
\begin{aligned}
(t \oplus r) &\mathbin{\widehat{\sqsubset}} (t \oplus s_1) && (6.59)\\
(t \oplus r) &\mathbin{\widehat{\sqsubset}} (t \oplus s_2) && (6.60)\\
(r \oplus t) &\mathbin{\widehat{\sqsubset}} (s_1 \oplus t) && (6.61)\\
(r \oplus t) &\mathbin{\widehat{\sqsubset}} (s_2 \oplus t)\,. && (6.62)
\end{aligned}
$$

Gelet op de definitie van zwakke deelstring en wegens $t \sqsubset t \oplus s_1$ en $t \sqsubset t \oplus s_2$, moet de langste deelstring van $t \oplus s_1$ en $t \oplus s_2$ van de vorm $t \oplus r$ zijn. Rekening houdend met de definitie van zwakke doorsnede moet r in deze vorm noodzakelijk voorkomen in $s_1 \mathbin{\widehat{\sqcap}} s_2$. Gelet op de definitie van zwakke deelstring en wegens $t \sqsubset s_1 \oplus t$ en $t \sqsubset s_2 \oplus t$, moet de langste deelstring van $s_1 \oplus t$ en $s_2 \oplus t$ van de vorm $r \oplus t$ zijn. Rekening houdend met de definitie van zwakke doorsnede moet r in deze vorm noodzakelijk voorkomen in $s_1 \mathbin{\widehat{\sqcap}} s_2$. □

Met betrekking tot de editeeroperaties van Levenshtein en Damerau geldt het volgende.

Eigenschap 6.3
Voor willekeurige strings $(s, t, a, b) \in \mathcal{S}^4$ en $c \in \mathcal{S}_1$ geldt er dat:

$$(s = a \oplus c \oplus b) \wedge (t = a \oplus b) \Rightarrow (s \mathbin{\widehat{\sqcap}} t = \{a \oplus b\}). \quad (6.63)$$

Bewijs. Het bewijs volgt onmiddellijk uit Eigenschap 6.2. □

Eigenschap 6.4
Voor willekeurige strings $(s,t,a,b) \in \mathcal{S}^4$ *en* $(c,d) \in (\mathcal{S}_1)^2$ *waarbij* $c \neq d$*, geldt er dat:*

$$(s = a \oplus c \oplus d \oplus b) \wedge (t = a \oplus d \oplus c \oplus b) \Rightarrow (s \,\widehat{\sqcap}\, t = \{a \oplus c \oplus b, a \oplus d \oplus b\}). \quad (6.64)$$

Bewijs. Het bewijs volgt onmiddellijk uit Eigenschap 6.2. □

Voor een verzameling van strings met een gelijke lengte definiëren we het concept karakterverzameling als volgt:

Definitie 6.11 (Karakterverzameling)
Voor een verzameling V *van strings met lengte* n $(V \subseteq \mathcal{S}_n)$ *is de* i^{de} *karakterverzameling van* V *bepaald als de deelverzameling van het alfabet* $\mathscr{A}$ *die de karakters bevat voorkomend op de* i^{de} *positie van strings in* V*:*

$$\mathscr{C}_{V,i} = \{c | c \in \mathscr{A} \wedge (\exists s \in V : \tau_s(i) = c)\}. \quad (6.65)$$

De globale karakterverzameling van V *wordt genoteerd als* $\mathscr{C}_V$ *en is gelijk aan:*

$$\mathscr{C}_V = \bigcup_{i=1}^{n} \mathscr{C}_{V,i}. \quad (6.66)$$

In een volgende sectie zullen de voorgaande definities worden gebruikt bij de constructie van evaluatoren voor strings.

6.4 Eén-niveau evaluatie voor strings

6.4.1 Definitie

In deze sectie willen we een nieuwe karaktergebaseerde evaluator voor strings construeren die louter op het niveau van karakters een vergelijkende analyse maakt van twee strings en vervolgens de mogelijkheid van coreferentie bepaalt. Deze evaluator wordt een één-niveau evaluator genoemd en zal in Sectie 6.5 worden gebruikt bij de constructie van een twee-niveau evaluator door toepassing van de resultaten van Hoofdstuk 4.

In onze aanpak voor karaktergebaseerde evaluatie besteden we aandacht aan twee zaken. Ten eerste blijkt uit Hoofdstuk 4 dat de complexiteit van een evaluator voor collecties afhangt van de complexiteit van de evaluator voor elementen. Dit wil zeggen dat de complexiteit van de één-niveau evaluator die in deze sectie wordt beschouwd, een belangrijke rol speelt bij het bepalen van de complexiteit van de evaluator die verderop wordt bestudeerd. De complexiteit van karaktergebaseerde methoden is meestal kwadratisch in termen van de lengtes van de strings die worden vergeleken. In onze aanpak proberen we dit te verbeteren. Ten tweede willen we in onze aanpak aandacht besteden aan het verschil tussen twee strings en meer bepaald aan de manier waarop het verschil zich verhoudt tot de oorspronkelijke strings. Voor twee strings s en t bestaat

er voor elke string r in de verzameling $s \mathbin{\widehat{\ominus}} t$ een vector $\mathbf{q}$ zodat $s[\mathbf{q}] = r$. De manier waarop deze vector $\mathbf{q}$ zich verhoudt tot de indexverzameling van s en t zal het criterium vormen voor de beslissing van coreferentie.

Teneinde de complexiteit van de evaluator beperkt te houden, zullen we een algoritme voorstellen dat een benadering maakt van $s \mathbin{\widehat{\sqcap}} t$ op zodanige wijze dat voor coreferente strings de benadering zeer sterk aansluit bij $s \mathbin{\widehat{\cap}} t$. Voor niet-coreferente strings zal onze benadering typisch minder goed aansluiten bij de werkelijke zwakke doorsnede. Het zal blijken dat het niet goed benaderen van de zwakke doorsnede in dit geval weinig tot geen impact heeft, gezien de invariabiliteit in de beslissing over coreferentie. Dit wil zeggen dat de beslissing van niet-coreferentie voor strings die inderdaad niet coreferent zijn, weinig wordt beïnvloed door het niet goed benaderen van de zwakke doorsnede. We illustreren eerst de basisidee achter de benadering. Voor de zwakke doorsnede geldt de volgende eigenschap.

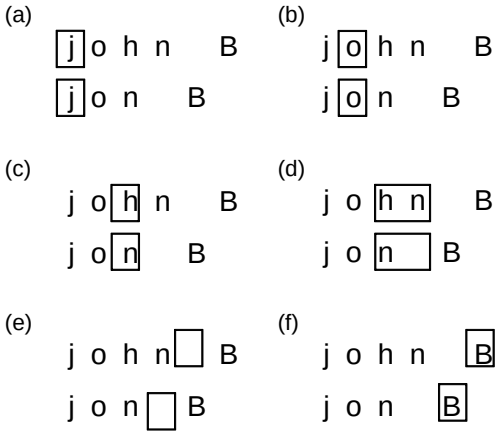
Eigenschap 6.5
Als voor de zwakke doorsnede van twee strings s en t geldt dat $\text{len}(s \mathbin{\widehat{\sqcap}} t) - 1$ *karakterverzamelingen singletons zijn, dan is de doorsnede volledig bepaald als het product van de karakterverzamelingen. Anders gezegd, als bijvoorbeeld de eerste* $\text{len}(s \mathbin{\widehat{\sqcap}} t) - 1$ *karakterverzamelingen singletons zijn, dan geldt er:*

$$\begin{aligned} \forall i \in \{1, ..., \text{len}(s \mathbin{\widehat{\sqcap}} t) - 1\} \quad & : |\mathscr{C}_{s \mathbin{\widehat{\cap}} t,i}| = 1 \\ & \Rightarrow s \mathbin{\widehat{\sqcap}} t = \mathscr{C}_{s \mathbin{\widehat{\cap}} t,1} \times ... \times \mathscr{C}_{s \mathbin{\widehat{\cap}} t,\text{len}(s \mathbin{\widehat{\cap}} t)}. \end{aligned} \tag{6.67}$$

Bewijs. Het bewijs volgt onmiddellijk uit de definitie van $\mathbin{\widehat{\sqsubset}}$ en $\mathbin{\widehat{\sqcap}}$. □

Als voldaan is aan de voorwaarde van Eigenschap 6.5, kan de zwakke doorsnede worden bepaald met behulp van zogenaamde *schuivende vensters*. Startend aan het begin van beide strings wordt een venster geschoven over elk van de strings. De grootte van deze vensters is initieel 1, kan incrementeel groeien en kan terugvallen naar 1. Karakters gedeeld door beide vensters worden toegevoegd aan de zwakke doorsnede. Veronderstel bijvoorbeeld de strings s="john B" en t="jon B". De constructie van de zwakke doorsnede $s \mathbin{\widehat{\sqcap}} t$ met behulp van schuivende vensters is geïllustreerd in Figuur 6.1.

In stappen (a) en (b) tonen de vensters voor beide strings respectievelijk de deelstrings "j" en "o". Deze beide deelstrings worden aan de zwakke doorsnede toegevoegd. In stap (c) tonen de vensters deelstrings in s en t die geen karakters gemeenschappelijk hebben. De venstergrootte wordt daarom verhoogd met 1. In stap (d) tonen de vensters deelstrings in s en t die het karakter 'n' gemeenschappelijk hebben. Dit karakter wordt toegevoegd aan de zwakke doorsnede. De venstergrootte valt terug naar de basisgrootte (d.i. 1) en de vensters schuiven op. Het opschuiven houdt in dat de vensters beginnen net na de positie waar het gemeenschappelijke karakter werd aangetroffen. In stap (d) worden de gemeenschappelijke karakters aangetroffen op posities 4 en 3, zodat de vensters in stap (e) zullen beginnen op posities 5 en 4. In stappen (e) en (f) tonen de vensters telkens identieke deelstrings die aan de zwakke doorsnede worden toegevoegd. Het resultaat is $s \mathbin{\widehat{\sqcap}} t = \{\text{"jon B"}\}$, wat inderdaad



Figuur 6.1: Venstergebaseerde constructie van de zwakke doorsnede van de strings "john B" en "jon B"

overeenkomt met de definitie van de zwakke doorsnede. Het kan worden ingezien dat zonder de veronderstelling van de singleton karakterverzamelingen, de zwakke doorsnede niet noodzakelijk correct wordt gevonden. Omwille van Eigenschappen 6.3 en 6.4 weten we dat als strings naar elkaar getransformeerd kunnen worden door middel van één editeeroperatie, dan is onze benadering correct. In deze thesis willen we aantonen dat de benadering met behulp van de *schuivende vensters* voldoende is om coreferente strings te kunnen detecteren. Uit de resultaten in Sectie 6.8 zal blijken dat onze evaluator een hoge *zuiverheid* heeft. We zullen aantonen dat de complexiteit van de methode met de vensters kwadratisch is in het slechtste geval, maar dat dit in de praktijk wordt vermeden.

Het zwak verschil tussen twee strings kan eveneens worden benaderd. Volgens Definitie 6.10 zijn deze verschillen verzamelingen van strings. We hebben al aangehaald dat voor twee strings s en t, elke string $r \in s \mathbin{\widehat{\ominus}} t$ kan worden geschreven als $s[\mathbf{q}]$ met $\mathbf{q}$ een transformatievector. In wat volgt zijn we geïnteresseerd in de deelvectoren van $\mathbf{q}$ die opeenvolgende indexen hebben.

Voorbeeld 6.7
Beschouw de volgende strings:

$$\begin{aligned} s &= \text{“st pietersnieuwstr”} \\ t &= \text{“sint pietersnieuwstraat”} \end{aligned}$$

dan vinden we dat $s \mathbin{\widehat{\sqcap}} t = \{\text{“st pietersnieuwstr”}\}$. *Het zwak verschil* $t \mathbin{\widehat{\ominus}} s$ *bevat precies één string, namelijk r=“inaat”. We kunnen nu zien dat er voor de transformatievector* $\mathbf{q} = (2, 3, 21, 22, 23)$ *geldt dat* $r = t[\mathbf{q}]$. *In deze vector kunnen twee deelvectoren worden gevonden die opeenvolgende indexen hebben, namelijk* $\mathbf{q}_1 = (2, 3)$ *en* $\mathbf{q}_2 = (21, 22, 23)$. *Wanneer we deze deelvectoren*

als transformatievectoren gebruiken vinden we dat $t[\mathbf{q}_1]$="in" en $t[\mathbf{q}_2]$="aat", hetgeen overeenkomt met twee gebruikte afkortingen.

Zoals uit Voorbeeld 6.7 blijkt, zijn we bij het bepalen van onzekerheid over coreferentie vooral geïnteresseerd in de lengtes van bepaalde deelstrings van het zwakke verschil. Bovendien willen we weten of een deelstring van het zwak verschil $s \mathbin{\widehat{\ominus}} t$ overeenkomt met een deelstring van het zwak verschil $t \mathbin{\widehat{\ominus}} s$. Daarom definiëren we het concept deelverschil als volgt.

Definitie 6.12 (Deelverschil tussen strings)
Beschouw twee strings s en t en beschouw hun zwakke doorsnede $s \mathbin{\widehat{\cap}} t$. Voor een willekeurige $r \in (s \mathbin{\widehat{\cap}} t)$ beschouwen we de vector $\mathbf{q}^{s,r}$ van lengte $|r|+1$ als volgt:

$$\forall i \in \{1,...,|r|+1\} : \mathbf{q}_i^{s,r} = \begin{cases} 0 & \textbf{\textit{als}} \quad i = 1 \\ \mathrm{ind}_{\mathbf{q}_{i-1}^{s,r}}(s, \tau_r(i-1)) & \textbf{\textit{als}} \quad i > 1 \end{cases} \tag{6.68}$$

Analoog beschouwen we de vector $\mathbf{q}^{t,r}$ van lengte $|r|+1$ als volgt:

$$\forall i \in \{1,...,|r|+1\} : \mathbf{q}_i^{t,r} = \begin{cases} 0 & \textbf{\textit{als}} \quad i = 1 \\ \mathrm{ind}_{\mathbf{q}_{i-1}^{t,r}}(t, \tau_r(i-1)) & \textbf{\textit{als}} \quad i > 1 \end{cases} \tag{6.69}$$

De multiverzameling van deelverschillen tussen twee strings s en t met betrekking tot $r \in (s \mathbin{\widehat{\cap}} t)$ bevat koppels $(\Delta_1, \Delta_2) \in \mathbb{N}^2$ waarvoor $\max(\Delta_1, \Delta_2) > 0$ en waarvoor er ofwel geldt dat:

$$\exists i \in \{2,...,|r|+1\} : \left(\Delta_1 = \mathbf{q}_i^{s,r} - \mathbf{q}_{i-1}^{s,r} - 1\right) \wedge \left(\Delta_2 = \mathbf{q}_i^{t,r} - \mathbf{q}_{i-1}^{t,r} - 1\right) \tag{6.70}$$

ofwel:

$$\left(\Delta_1 = |s| - \mathbf{q}_{|r|+1}^{s,r}\right) \wedge \left(\Delta_2 = |t| - \mathbf{q}_{|r|+1}^{t,r}\right). \tag{6.71}$$

Voor een deelverschil $\Delta = (\Delta_1, \Delta_2)$ wordt $\max(\Delta_1, \Delta_2)$ de lengte van het deelverschil genoemd.

Voorbeeld 6.8
Beschouw de strings s="st pieter" en t="sint pieter", dan geldt er dat:

$$s \mathbin{\widehat{\cap}} t = \{\text{"st pieter"}\}. \tag{6.72}$$

Voor r="st pieter" vinden we dat:

$$\mathbf{q}^{s,r} = (0,1,2,3,4,5,6,7,8,9) \tag{6.73}$$
$$\mathbf{q}^{t,r} = (0,1,4,5,6,7,8,9,10,11). \tag{6.74}$$

We zien dan voor $i = 3$ dat:

$$\Delta_1 = \mathbf{q}_i^{s,r} - \mathbf{q}_{i-1}^{s,r} - 1 = 2 - 1 - 1 = 0 \tag{6.75}$$
$$\Delta_2 = \mathbf{q}_i^{t,r} - \mathbf{q}_{i-1}^{t,r} - 1 = 4 - 1 - 1 = 2. \tag{6.76}$$

Dit deelverschil $(0,2)$ is meteen het enige deelverschil dat we hier vaststellen.

In wat volgt zullen we vier types van deelverschillen tussen strings beschouwen. Deze types worden getoond in Tabel 6.1. Wanneer een deelverschil Δ wordt gevonden voor $i = 2$ en $\min(\Delta_1, \Delta_2) = 0$ spreekt men van een prefix. Wanneer een deelverschil Δ voortkomt uit (6.71) en $\min(\Delta_1, \Delta_2) = 0$ spreekt men van een suffix. In elk ander geval waarbij $\min(\Delta_1, \Delta_2) = 0$ spreekt men van een gaping en in elk geval waarbij $\min(\Delta_1, \Delta_2) \neq 0$ spreekt men van een fout. Tijdens de constructie van de zwakke doorsnede $s \mathbin{\widehat{\cap}} t$ kan worden geverifieerd welke van deze types wordt aangetroffen, zodat een beslissing over coreferentie genomen kan worden op basis van de aangetroffen deelverschillen.

Type	s	t	(Δ_1, Δ_2)
prefix	koud	**ij**skoud	(0,3)
suffix	str	str**aat**	(0,3)
gaping	jon	jo**h**n	(0,1)
fout	gr**oe**n	gr**aa**n	(2,2)

Tabel 6.1: Deelverschillen tussen strings

We geven nu eerst de definitie van onze evaluator en leggen het bijhorende algoritme uit.

Definitie 6.13 (Eén-niveau evaluator voor strings)
Een één-niveau evaluator voor strings is een functie $E_{\mathcal{S}}$ gedefinieerd als:

$$E_{\mathcal{S}} : \mathcal{S}^2 \to \mathcal{F}(\mathbb{B}) : (s, t) \mapsto E_{\mathcal{S}}(s, t) \tag{6.77}$$

De manier waarop een evaluator $E_{\mathcal{S}}$ de onzekerheid over coreferentie bepaalt, is gebaseerd op de eerder vermelde benadering van de zwakke doorsnede van twee strings en de bijhorende deelverschillen. Onze aanpak is vastgelegd in Algoritme 6.1. Het algoritme begint met de initialisatie van een aantal parameters (regel 1). Zoals reeds is aangehaald willen we een venster laten schuiven over elk van de strings. Deze vensters komen telkens overeen met deelstrings van s en t. De wijzers w_s en w_t geven telkens de eerste index aan van de deelstring die overeenkomt met het venster, terwijl v de verschuiving aangeeft. De grootte van het venster is steeds gelijk aan de grootte van de verschuiving vermeerderd met 1. De combinatie van een wijzer met de verschuiving bepaalt steeds de deelstring die binnen het venster valt. Het algoritme werkt stapsgewijs en probeert bij elke stap één karakter van elke string toe te voegen aan het venster voor die string (regel 4-12). De Boolese variabele f_s (resp. f_t) geeft aan wanneer deze toevoeging niet meer mogelijk is voor s (resp. t). Als de karakterverzamelingen van de deelstrings overeenkomend met de vensters disjunct zijn, wordt de verschuiving verhoogd met 1 en wordt overgegaan naar de volgende stap (regel 20-21). Zoniet, dan betekent dit dat de strings s en t een karakter gemeenschappelijk hebben in hun vensters. De lengte van de doorsnede, bijgehouden in variabele x wordt dan met één verhoogd. Als $v > 0$, dan wordt er een deelverschil toegevoegd aan de lijst L_Δ die alle deelverschillen bijhoudt (regel 22-30). In stap (d) van Figuur 6.1 is bijvoorbeeld te zien

Algoritme 6.1 $E_{\mathcal{S}}(s,t)$

1: $(next, w_s, w_t, v, x) \leftarrow (\textbf{true}, 1, 1, 0, 0)$
2: $(f_s, f_t) \leftarrow (\textbf{false}, \textbf{false})$
3: **repeat**
4: **if** $w_s + v \leq |s|$ **then**
5: $a \leftarrow \tau_s(w_s + v)$
6: **else**
7: $(f_s, f_t) \leftarrow (\textbf{true}, f_t \vee (w_s > |s|))$
8: **end if**
9: **if** $w_t + v \leq |t|$ **then**
10: $b \leftarrow \tau_t(w_t + v)$
11: **else**
12: $(f_s, f_t) \leftarrow (f_s \vee (w_t > |t|), \textbf{true})$
13: **end if**
14: **if** $f_s \wedge f_t$ **then**
15: $next \leftarrow \textbf{false}$
16: $(L_\Delta).\textbf{add}(|s| - w_s + v + 1, |t| - w_t + v + 1, w_s, w_t)$
17: **else**
18: $ind_s \leftarrow \text{ind}(s[(w_s, ..., \min(|s|, w_s + v))], b)$
19: $ind_t \leftarrow \text{ind}(t[(w_t, ..., \min(|t|, w_t + v))], a)$
20: **if** $ind_s = 0 \wedge ind_t = 0$ **then**
21: $v \leftarrow v + 1$
22: **else if** $(\min(ind_s, ind_t) > 0 \wedge ind_s < ind_t) \vee ind_t = 0$ **then**
23: $(L_\Delta).\textbf{add}(ind_s - 1, v + 1, w_s, w_t)$
24: $(x, w_s, w_t, v) \leftarrow (x + 1, w_s + ind_s, w_t + v + 1, 0)$
25: **else if** $(\min(ind_s, ind_t) > 0 \wedge ind_t < ind_s) \vee ind_s = 0$ **then**
26: $(L_\Delta).\textbf{add}(v + 1, ind_t - 1, w_s, w_t)$
27: $(x, w_s, w_t, v) \leftarrow (x + 1, w_s + v + 1, w_t + ind_t, 0)$
28: **else**
29: $(L_\Delta).\textbf{add}(v, v, w_s, w_t)$
30: $(x, w_s, w_t, v) \leftarrow (x + 1, w_s + v + 1, w_t + v + 1, 0)$
31: **end if**
32: **end if**
33: **until** $next = \textbf{false}$
34: **return** $\textbf{uncertainty}(x, L_\Delta)$

hoe de vensters de deelstrings “hn” in s en “n ” in t omvatten. Het gemeenschappelijke karakter ‘n’ wordt toegevoegd aan de zwakke doorsnede en een deelverschil $(1, 0)$ wordt toegevoegd aan L_Δ. Op basis van de wijzers w_s en w_t weten we dat het deelverschil van het type ‘gaping’ is. Stap voor stap wordt de lijst met deelverschillen aangevuld tot alle karakters onderzocht zijn. Daarna wordt op basis van de lijst met deelverschillen een possibilistische waarheidswaarde opgesteld die de mogelijkheid weergeeft dat s en t (niet) coreferent zijn (regel 34). Laat ons bij wijze van voorbeeld opnieuw de strings s=“john B”

en t="jon B" beschouwen. Figuur 6.2 illustreert de werking van Algoritme 6.1 door na elke iteratie de waarde van de belangrijkste parameters te tonen. Hierbij wordt $(w_s, ..., \min(|s|, w_s + v))$ kort genoteerd als $\mathbf{q}_s$ = (regel 18) en wordt $(w_t, ..., \min(|t|, w_t + v))$ kort genoteerd als $\mathbf{q}_t$ (regel 19).

	w_s	w_t	v	next	a	b	$s[\mathbf{q}_s]$	$t[\mathbf{q}_t]$	x	ind_s	ind_t	L_Δ
init	0	0	0	true	''	''	σ	σ	0	0	0	{}
1	1	1	0	true	'j'	'j'	"j"	"j"	1	1	1	{}
2	3	3	0	true	'o'	'o'	"o"	"o"	2	1	1	{}
3	3	3	1	true	'h'	'n'	"h"	"n"	2	0	0	{}
4	5	4	0	true	'n'	' '	"hn"	"n "	3	0	1	{(0,1)}
5	6	5	0	true	' '	' '	" "	" "	4	1	1	{(0,1)}
6	7	6	0	true	'B'	'B'	"B"	"B"	5	1	1	{(0,1)}
7	-	-	-	false	-	-	-	-	-	-	-	{(0,1)}

Figuur 6.2: Toepassing van Algoritme 6.1 op de strings "john B" en "jon B"

Het is duidelijk dat de manier waarop deelverschillen worden beoordeeld van cruciaal belang is voor het verkrijgen van een goede evaluator. We willen dit aspect enigzins flexibel houden omdat het gebruikte toekenningsmodel voor possibiliteit wordt bepaald door de manier waarop evaluator $E_{\mathcal{S}}$ wordt gebruikt. Wanneer $E_{\mathcal{S}}$ bijvoorbeeld wordt gebruikt door een twee-niveau evaluator kan het toekenningsmodel veel strenger zijn dan wanneer $E_{\mathcal{S}}$ als alleenstaande evaluator optreedt (Sectie 6.8). Het toekenningsmodel dat in deze thesis wordt vooropgesteld, beschikt daarom voor elk type deelverschil over een lengtegrens ($\in \mathbb{N}$) en een kost ($\in \mathbb{R}$). Als in de lijst van deelverschillen L_Δ een deelverschil voorkomt dat groter is dan de toegestane lengte voor dat type, dan wordt beslist dat de strings zeker niet coreferent zijn en wordt de possibilistische waarheidswaarde $(0, 1)$ gegeven als resultaat van de vergelijking. Anders gezegd:

$$\exists \Delta \in L_\Delta : \max(\Delta_1, \Delta_2) > \text{grens}(\text{type}(\Delta)) : E_{\mathcal{S}}(s, t) = (0, 1). \tag{6.78}$$

Merk op dat het type deelverschil in Algoritme 6.1 kan worden afgeleid uit de combinatie van de verschuiving v en de wijzers w_s en w_t. Als alle deelverschillen binnen de toegestane grenzen vallen, wordt enerzijds de totale kost gezien als een evidentie van niet-coreferentie en wordt deze evidentie berekend als volgt:

$$e_F = \sum_{\Delta \in L_\Delta} \text{kost}\left(\text{type}\left(\Delta\right)\right). \tag{6.79}$$

Anderzijds wordt de lengte van de doorsnede gedeeld door het maximum van de stringlengtes, hetgeen gezien kan worden als een evidentie voor coreferentie.

$$e_T = \frac{\text{len}(s \mathbin{\widehat{\sqcap}} t)}{\max(|s|, |t|)}. \tag{6.80}$$

Het resultaat is dan een genormaliseerde possibilistische waarheidswaarde die wordt berekend als:

$$\widetilde{p} = \left(\frac{e_T}{\max(e_T, e_F)}, \frac{e_F}{\max(e_T, e_F)} \right). \tag{6.81}$$

In wat volgt beschouwen we drie instanties van dit toekenningsmodel. Het eerste model wordt getoond in Tabel 6.2 en geeft aanleiding tot een zogenoemde suffixevaluator, waarbij $c \in]0, 1]$ een constante is. In dit toekenningsmodel zijn strings enkel coreferent als het enige deelverschil een suffix is. De lengte van deze suffix wordt niet begrensd. Een suffixevaluator is symmetrisch en sterk

Type	grens	kost
prefix	0	-
suffix	∞	c
gaping	0	-
fout	0	-

Tabel 6.2: Toekenningsmodel voor een suffixevaluator

reflexief, aangezien $c \neq 0$. Het kan worden ingezien dat de benadering van de zwakke doorsnede door Algoritme 6.1 altijd correct is voor strings die enkel in een suffix verschillen. Een suffixevaluator is niet noodzakelijk transitief en niet noodzakelijk consistent. Een tweede toekenningsmodel wordt getoond in Tabel 6.3 en geeft aanleiding tot een zogenoemde strenge evaluator. Dit model laat een deelverschil van het type suffix toe en tolereert gapingen met een lengte kleiner dan 3. Een dergelijke strenge evaluator wordt in wat volgt gebruikt door een twee-niveau evaluator om deelstrings te vergelijken. Een strenge evaluator is steeds symmetrisch en sterk reflexief.

Type	grens	kost
prefix	0	-
suffix	∞	0.1
gaping	2	0.3
fout	0	-

Tabel 6.3: Toekenningsmodel voor een strenge evaluator

Het derde en laatste toekenningsmodel is getoond in Tabel 6.4 en geeft aanleiding tot een zogenoemde standaardevaluator. Dit model laat alle types tot op zekere hoogte toe. Een standaardevaluator zal worden gebruikt wanneer $E_{\mathcal{S}}$ als een alleenstaande evaluator optreedt, zonder deel uit te maken van

een twee-niveau evaluator. Een standaardevaluator is steeds symmetrisch en sterk reflexief. De kosten en lengtegrenzen zijn empirisch bepaald en komen overeen met een standaard foutenmodel dat rekening houdt met schrijffouten en afkortingen, zij het in beperkte mate zoals blijkt uit onze experimenten (Sectie 6.8).

Type	grens	kost
prefix	10	0.2
suffix	10	0.1
gaping	2	0.2
fout	2	0.3

Tabel 6.4: Toekenningsmodel voor een standaardevaluator

6.4.2 Complexiteitsanalyse

Zoals eerder is aangehaald, wordt de lengte van de zwakke doorsnede slechts bij benadering berekend. Het voordeel dat we hier willen uithalen is een efficiënte berekening van de benadering. Om die reden maken we hier een analyse van de complexiteit van een één-niveau evaluator. Deze complexiteit hangt af van het aantal vergelijkingen van karakters dat nodig is tijdens het uitvoeren van het algoritme. We veronderstellen tijdens deze analyse dat $|s| \leq |t|$, zonder verlies van de algemeenheid.

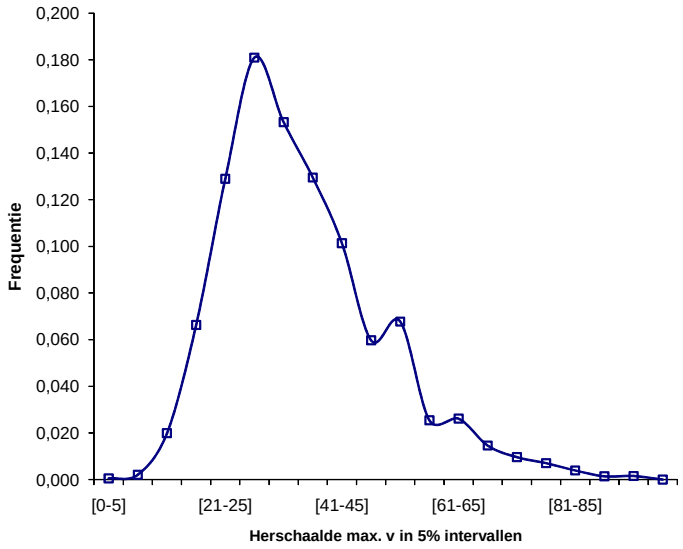
De complexiteit van één iteratie hangt af van de berekening van ind_s en ind_t (regels 18-19 van Algoritme 6.1). Als in iteratie i zowel ind_s als ind_s gelijk zijn aan 0, dan wil dit zeggen dat de karakterverzamelingen van de deelstrings aangeduid door de vensters disjunct zijn. De verschuiving en bijgevolg dus ook de venstergrootte verhoogt dan met 1. De wijzers w_s en w_t blijven dezelfde. Bijgevolg zijn beide vensters in iteratie $i+1$ een eenheid groter dan in iteratie i. Bemerk dat beide vensters steeds dezelfde grootte $v+1$ hebben. Noteren we de deelstrings aangeduid door de vensters in s en t met b_s en b_t (regels 18-19 van Algoritme 6.1), dan geldt:

$$\mathscr{C}_{b_s[(1,\ldots,|b_s|-1)]} \cap \mathscr{C}_{b_t[(1,\ldots,|b_t|-1)]} = \emptyset. \tag{6.82}$$

Bij elke iteratie wordt het laatste karakter van zowel b_s als b_t beschouwd (respectievelijk variabelen a en b in Algoritme 6.1) en wordt a in b_t en b in b_s gezocht. Bijgevolg kan het aantal karaktervergelijkingen (vgln_v) nodig in iteratie i worden uitgedrukt in termen van de verschuiving v. Als $w_s + v < |s|$ dan vinden we:

$$\text{vgln}_v = \begin{cases} 1 & v = 0 \\ 2(v+1) & v > 0. \end{cases} \tag{6.83}$$

In het geval waar $w_s + v \geq |s|$ is het aantal karaktervergelijkingen in een iteratie $|s| - w_s$. Doorheen verschillende iteraties kan de grootte van het venster



Figuur 6.3: Frequenties van de verhoudingen van maximale venstergrootte over maximale stringlengte

lineair toenemen en hoe groter de venstergrootte, hoe meer karaktervergelijkingen gedaan moeten worden. Gelet op (6.83) geldt er dat als $v < |s|$, het aantal karaktervergelijkingen sinds de laatste keer dat $v = 0$ kan worden geschreven als:

$$\sum_{i=0}^{v} \text{vgln}_i = 1 + 2\sum_{i=1}^{v}(i+1) = 1 + \frac{2v(v+1)}{2} + 2v = v^2 + 3v + 1 \qquad (6.84)$$

waaruit volgt dat het aantal karaktervergelijkingen nodig om een nieuw karakter toe te voegen aan de zwakke doorsnede een kwadratische complexiteit heeft ($O(v^2)$). Bijgevolg is de complexiteit van $E_{\mathcal{S}}$ $O(v^2)$ in plaats van $O(|s||t|)$. De kracht van het algoritme ligt in het feit dat een kleine venstergrootte een lagere complexiteit met zich meebrengt. We kunnen nu empirisch gaan verifiëren dat de gemiddelde venstergrootte van het algoritme relatief laag is. Figuur 6.3 toont de frequentie van de verhoudingen $\frac{\max v}{\max(|s|,|t|)}$ van de maximale venstergrootte bereikt tijdens de vergelijking van twee strings tot $\max(|s|, |t|)$. Deze verhoudingen $\frac{\max v}{\max(|s|,|t|)}$ worden berekend voor elke stringvergelijking en samengenomen in intervallen van 5%, teneinde de grafiek uit te middelen. Figuur 6.3 is het resultaat van 852930 stringvergelijkingen in een representatieve datacollectie. Het kan worden vastgesteld dat de gemiddelde maximale venstergrootte ongeveer 30% van de lengte van de langste string onder vergelijking is. In Sectie 6.8 zullen we de uitvoeringstijden van verschillende karaktergebaseerde methoden vergelijken en zullen we de uitvoeringstijd bestuderen in functie van de gemiddelde stringlengte.

6.5 Twee-niveau evaluatie voor strings

In deze sectie zullen we bestuderen hoe we een twee-niveau evaluator voor strings kunnen construeren door gebruik te maken van de resultaten uit Hoofdstuk 4 en waarbij een één-niveau evaluator uit vorige sectie dienst doet als evaluator op het laagste niveau (d.i. het niveau van elementen). Om het principe van evaluatie van collecties te kunnen toepassen, wordt gebruikt gemaakt van een splitsingsfunctie $\mathscr{S}$ om een string s te transformeren naar een multiverzameling van deelstrings. In de meeste gevallen komen deze deelstrings overeen met woorden. Zoals is aangehaald in Hoofdstuk 2 kan een dergelijke transformatiefunctie worden gezien als een uitbreiding van het meetproces. Aangezien meetprocessen imperfect kunnen zijn, betekent dit dat ook transformatiefuncties imperfect kunnen zijn. Dergelijke imperfecties zullen hier worden onderzocht. Het nut van twee-niveau systemen komt voort uit het feit dat karaktergebaseerde systemen vaak ontoereikend zijn voor strings met een hogere gemiddelde lengte. In onze aanpak veronderstellen we dat het resultaat van een splitsingsfunctie een (multi)verzameling is. De ordening van de deelstrings wordt bijgevolg niet bewaard. Deze beslissing is genomen vermits bestaande methoden hebben aangetoond dat deze ordening vaak niet belangrijk is voor het vergelijken van strings [90, 91]. Alle bestaande twee-niveau systemen, die gebruik maken van een flexibele vergelijking van deelstrings zijn steeds gebaseerd op een vectorruimtemodel, waarbij het TFIDF gewichtsschema gebruikt wordt om de vectoren te construeren en waarbij similariteit tussen vectoren wordt berekend als de cosinus van de hoek tussen twee vectoren. In deze thesis willen we een nieuw twee-niveau systeem voorstellen, met een aantal opmerkelijke voordelen ten opzichte van de bestaande methoden. Laat ons beginnen met een definitie van een splitsingsfunctie.

Definitie 6.14 (Splitsingsfunctie)
Gegeven een alfabet $\mathscr{A}$ en een niet-lege deelverzameling van dit alfabet $\mathscr{G} \subset \mathscr{A}$ die de grensverzameling wordt genoemd. Een splitsingsfunctie $\mathscr{S}_{\mathscr{G}}$ ten opzichte van de grensverzameling $\mathscr{G}$ is een functie:

$$\mathscr{S}_{\mathscr{G}} : \mathcal{S} \to \mathcal{M}(\mathcal{S}) \tag{6.85}$$

zodat voor een string s de multiverzameling $\mathscr{S}_{\mathscr{G}}(s)$ gelijk is aan de multiverzameling van deelstrings van s waarvoor geldt:

$$\forall t \in \mathscr{S}_{\mathscr{G}}(s) : \exists \mathbf{q} \in \mathbb{N}^{|t|} : (s'[\mathbf{q}] = t) \wedge (\tau_{s'}(\mathbf{q}_1 - 1) \in \mathscr{G}) \wedge (\tau_{s'}(\mathbf{q}_{|t|} + 1) \in \mathscr{G}). \tag{6.86}$$

Hierbij is s' een uitbreiding van s door vooraan en achteraan een karakter uit de grensverzameling te concateneren. Dit betekent dat er voor s' moet gelden dat:

$$\exists r \in \mathcal{S}_1 : (\tau_r(0) \in \mathscr{G}) \wedge (s' = r \oplus s \oplus r)\,. \tag{6.87}$$

Definitie 6.14 stelt dat splitsen van een string equivalent is aan het verwijderen van alle karakters die voorkomen in de grensverzameling. Deze grensverzameling bevat naast het blanco karakter (d.i. witruimte) typisch ook leestekens.

Het splitsen heeft bijgevolg een lineaire complexiteit ($O(|s|)$). Merk op dat het resultaat van een splitsing inderdaad een multiverzameling is, aangezien eenzelfde deelstring meerdere keren kan voorkomen binnen een string.

Voorbeeld 6.9
Beschouw de string s="De Muur (Muur van Geraardsbergen)". Veronderstel $\mathscr{G} = \{$'(',')',' '$\}$. *Het toepassen van de splitsingsfunctie levert een multiverzameling:*

$$M = \mathscr{S}_{\mathscr{G}}(s) \tag{6.88}$$

die wordt gekarakteriseerd door de volgende multipliciteitsfunctie:

$$\begin{aligned} \omega_M(\text{“Muur”}) &= 2 && (6.89)\\ \omega_M(\text{“De”}) &= 1 && (6.90)\\ \omega_M(\text{“van”}) &= 1 && (6.91)\\ \omega_M(\text{“Geraardsbergen”}) &= 1. && (6.92) \end{aligned}$$

Definitie 6.15 (Twee-niveau evaluator voor strings)
Een twee-niveau evaluator voor strings is gedefinieerd als:

$$E^*_{\mathcal{S}} : \mathcal{S}^2 \to \mathcal{F}(\mathbb{B}) : (s,t) \mapsto E_{\mathcal{M}(\mathcal{S})}(\mathscr{S}_{\mathscr{G}}(s), \mathscr{S}_{\mathscr{G}}(t)). \tag{6.93}$$

Definitie 6.15 stelt dat een twee-niveau evaluator $E^*_{\mathcal{S}}$ twee strings vergelijkt door ze op te splitsen in deelstrings en vervolgens de verkregen collecties van deelstrings te vergelijken. Hiervoor wordt een kwantorfunctie gebruikt die uitdrukt hoeveel deelstrings coreferent moeten zijn alvorens besloten kan worden dat de strings coreferent zijn (Hoofdstuk 4).

De eigenschappen van een twee-niveau evaluator zijn de volgende. Een twee-niveau evaluator is symmetrisch en reflexief. Aangezien $\mathscr{S}_{\mathscr{G}}$ niet-injectief is, geldt er dat $E^*_{\mathcal{S}}$ niet sterk reflexief is. Deze eigenschap is belangrijk bij het recursieve gebruik van twee-niveau evaluatoren, aangezien een evaluator $E_{\mathcal{M}(U)}$ veronderstelt dat de evaluator E_U die wordt gebruikt sterk reflexief is. We veronderstellen echter steeds dat een twee-niveau evaluator $E^*_{\mathcal{S}}$ collecties van deelstrings vergelijkt op basis van een één-niveau evaluator die wel sterk reflexief is.

Door gebruik te maken van een kwantorfunctie (Hoofdstuk 4) voor het bepalen van conditionele necessiteit heeft het berekenen van de gewichten die de conditionele necessiteit voorstellen, een lage complexiteit. Bij technieken gebaseerd op het TFIDF schema moet het gewicht van elke deelstring apart worden berekend op basis van frequenties. Het aantal verschillende deelstrings in een datacollectie kan echter hoog oplopen, waardoor het opzoeken van deelstrings in een frequentietabel een significante complexiteit heeft. Deze complexiteit is niet aanwezig in onze aanpak en wordt gezien als een voordeel ten opzichte van bestaande technieken.

Kleine verschillen in de parameters van een kwantorfunctie zullen zwaarder doorwegen naarmate $|\mathscr{S}_{\mathscr{G}}(s)|$ kleiner wordt. Het resultaat van $E^*_{\mathcal{S}}$ is namelijk

gebaseerd op het aantal coreferente deelstrings. De bijdrage van één koppel deelstrings tot het totaal is:

$$\frac{1}{\max(|\mathscr{S}_{\mathscr{G}}(s)|, |\mathscr{S}_{\mathscr{G}}(t)|)}. \tag{6.94}$$

Hoe groter de gemiddelden van $|\mathscr{S}_{\mathscr{G}}(s)|$ en $|\mathscr{S}_{\mathscr{G}}(t)|$ worden voor een datacollectie, hoe kleiner de individuele bijdrage van een koppel deelstrings. Dit is een belangrijk inzicht aangezien we een strenge evaluator (Tabel 6.3) als één-niveau evaluator zullen veronderstellen, tenzij expliciet anders vermeld. Door dit strenge toekenningsmodel zal de één-niveau evaluator $E_{\mathcal{S}}$ slechts in een beperkt aantal gevallen een possibilistische waarheidswaarde verschillend van $(0, 1)$ toekennen. Het is hierdoor niet ondenkbaar dat $E_{\mathcal{S}}$ deze possibilistische waarheidswaarde ook toekent aan een koppel van coreferente deelstrings. Wat we in dit hoofdstuk ondermeer willen aantonen, is dat dergelijke foute toekenningen niet noodzakelijk een negatieve invloed hebben op de prestatie van de twee-niveau evaluator $E^*_{\mathcal{S}}$. Deze stelling verdedigen we door te wijzen op het feit dat dergelijke fouten niet worden gemaakt voor het merendeel van de koppels van deelstrings. De vaststelling die we hier maken met betrekking tot de bijdrage van één koppel deelstrings laat blijken dat de impact van foute toekenningen kleiner wordt naarmate het gemiddeld aantal deelstrings toeneemt.

Voorbeeld 6.10
Beschouw twee strings s=“beach palace” en t=“beach palace hotel”. We veronderstellen $\mathscr{G} = \{\text{‘ ’}\}$ zodat er geldt:

$$\mathscr{S}_{\mathscr{G}}(s) = \{\text{“beach”}, \text{“palace”}\} \tag{6.95}$$
$$\mathscr{S}_{\mathscr{G}}(t) = \{\text{“beach”}, \text{“palace”}, \text{“hotel”}\}. \tag{6.96}$$

Vervolgens veronderstellen we dat $E_{\mathcal{S}}$ gebruik maakt van het strenge toekenningsmodel. Dit geeft aanleiding tot de volgende matrix van possibilistische waarheidswaarden:

	“beach”	“palace”	“hotel”
“beach”	$(1, 0)$	$(0, 1)$	$(0, 1)$
“palace”	$(0, 1)$	$(1, 0)$	$(0, 1)$

(6.97)

Op basis van deze matrix wordt een injectieve afbeelding ι geconstrueerd tussen $\mathscr{S}_{\mathscr{G}}(s)$ en $\mathscr{S}_{\mathscr{G}}(t)$ (Hoofdstuk 4):

$$\begin{aligned} \iota(\text{“beach”}) &= \text{“beach”} \\ \iota(\text{“palace”}) &= \text{“palace”}. \end{aligned} \tag{6.98}$$

Gelet op deze afbeelding en rekening houdend met het verschil in kardinaliteit tussen $\mathscr{S}_{\mathscr{G}}(s)$ en $\mathscr{S}_{\mathscr{G}}(t)$ komen we tot de volgende collectie van possibilistische waarheidswaarden:

$$\{(1, 0), (1, 0), (0, 1)\}. \tag{6.99}$$

Combinatie van deze possibilistische waarheidswaarden (Hoofdstuk 3) levert ons de onzekerheid over coreferentie van s en t. Hierbij maken we gebruik van een kwantorfunctie voor het bepalen van conditionele necessiteit. Deze kwantorfunctie maakt gebruik van de parameters (α, β, δ). Stel bijvoorbeeld dat we als waarden voor deze parameters $(1, 0, 0)$ kiezen, dan verkrijgen we de functie getoond in Figuur 4.8. Onder deze veronderstelling vinden we dat $E^*_{\mathcal{S}}(s, t) = (1, 0)$.

6.6 Geavanceerde aspecten

In deze sectie willen we aandacht besteden aan een drietal aspecten die de evaluator uit Definitie 6.15 uitbreiden tot een meer geavanceerde techniek.

6.6.1 Splitsingsfunctie

Het eerste aspect handelt over de splitsingsfunctie. Deze functie kan worden gezien als een deel van het meetproces. Aangezien meetprocessen onderhevig kunnen zijn aan imperfecties, stellen we ons hier de vraag hoe eventuele imperfecties van de splitsingsfunctie een twee-niveau evaluator beïnvloeden. Vergelijking van collecties is gebaseerd op een injectieve afbeelding tussen twee collecties. Bij het splitsen van strings kan daardoor het volgende probleem ontstaan. Beschouw twee strings s="groteweg" en t="grote weg" en beschouw de grensverzameling $\mathscr{G} = \{\text{`}\ \text{'}\}$ ($\mathscr{G}$ bevat enkel het blanco karakter). Het kan worden ingezien dat s wordt opgesplitst in één deelstring en t in twee deelstrings. De splitsingsfunctie veroorzaakt hier een verschil in kardinaliteit tussen de verkregen collecties. Een mogelijke oplossing voor dit probleem is nagaan of concatenaties van elementen uit $\mathscr{S}_{\mathscr{G}}(s)$ voorkomen in $\mathscr{S}_{\mathscr{G}}(t)$ en omgekeerd. Uiteraard levert deze stap een bijkomende kwadratische complexiteit op. Een alternatieve oplossing is het gebruik van een geavanceerde splitsingsfunctie die met behulp van (grammaticale) taalregels een woord in twee of meer deelstrings kan opsplitsen. Dergelijke splitsingsfuncties worden typisch gebruikt door tekstverwerkers, maar zouden in de context van stringvergelijking een nuttige bijdrage kunnen leveren.

6.6.2 Frequentiefilter

Een tweede aspect dat we willen bespreken is het gebruik van een frequentiefilter. Door het gebruik van een kwantorfunctie wordt een voordeel verkregen op het vlak van complexiteit ten opzichte van andere methoden. Dit komt voort uit het feit dat per deelstring geen apart gewicht moet worden opgezocht. Echter, een voordeel van een gewichtsschema is dat bepaalde deelstrings als onbelangrijk kunnen worden bestempeld. Meestal wordt hiervoor de regel gebruikt dat deelstrings die vaak voorkomen in verschillende strings een laag gewicht krijgen. Dergelijke hoogfrequente deelstrings kunnen ook in onze possibilistische aanpak een storende rol spelen. Wanneer een deelstring r vaak

voorkomt in verschillende (niet-coreferente) strings, dan zullen heel wat koppels (s,t) bestaan waarvoor $\mathscr{S}_{\mathscr{G}}(s) \cap \mathscr{S}_{\mathscr{G}}(t)$ niet leeg is. Dit verschijnsel kan tot gevolg hebben dat de voorwaarde die wordt gesteld door de kwantorfunctie voor heel wat koppels a priori gedeeltelijk is voldaan. Als gevolg hiervan verliest de kwantorfunctie aan discriminerend vermogen. Willen we dit effect wegnemen, dan kunnen we geen gebruik maken van gewichten om verschillende redenen. Ten eerste is aangehaald in Hoofdstuk 4 dat het gebruik van een kwantorfunctie de enige garantie is op een unieke uitkomst van de vergelijking van twee collecties. Als we de conditionele necessiteit laten afhangen van de deelstrings die op elkaar worden afgebeeld onder ι, dan moet rekening worden gehouden met het feit dat ι niet uniek is. Ten tweede willen het voordeel in complexiteit, bereikt door gebruik van een kwantorfunctie, behouden. Ten derde is de impact van gewichten naar onze mening niet altijd duidelijk. In een publicatiedatabank zal de deelstring "2000" bijvoorbeeld veel voorkomen als er veel publicaties uit het jaar 2000 aanwezig zijn. Dit wil echter niet zeggen dat deze deelstring als onbelangrijk bestempeld moet worden. In tegendeel: het jaartal van een publicatie is discriminerende informatie. Om deze redenen willen we een alternatieve aanpak voorstellen die deelstrings wegfiltert als aan welbepaalde voorwaarden is voldaan. Deze filter zal worden toegepast na splitsing maar voor de constructie van de afbeelding tussen de collecties van deelstrings. Op die manier blijft het resultaat van de evaluatie uniek. Aangezien het filteren onafhankelijk van de vergelijking gebeurt, kan het in lineaire tijd $O(|A| + |B|)$ worden uitgevoerd, waarbij $A \times B$ de verzameling van koppels van strings voorstelt. Voor de filter kan een lijst met stopwoorden worden gebruikt [98], maar dat maakt het geheel contextafhankelijk. Daarom gebruiken we hier een filter gebaseerd op twee karakteristieken van een collectie van strings A: het aantal strings ($|A|$) en het aantal unieke deelstrings (u). Als u relatief laag is in vergelijking met $|A|$, dan betekent dit dat vele deelstrings worden herhaald in meerdere strings. Als u relatief hoog is in vergelijking met $|A|$, dan betekent dit dat de meeste deelstrings slechts voorkomen in een klein aantal strings. Om deze redenen baseren we de filter op de verhouding:

$$\frac{|A|}{u}. \tag{6.100}$$

Een deelstring r wordt weggefilterd als:

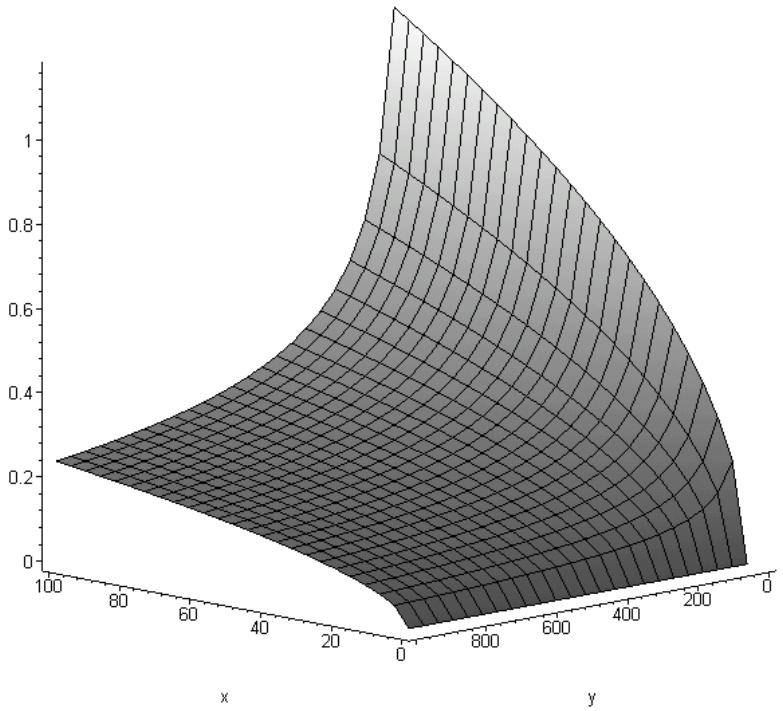
$$\text{aantal}(r) \geq \left(\sqrt{\frac{|A|}{u}}\frac{3}{4}\right)(|A|) \tag{6.101}$$

waarbij aantal(r) gelijk is aan het aantal strings s waarvoor $r \in \mathscr{S}_{\mathscr{G}}(s)$. Onder deze voorwaarden is de drempelwaarde voor de frequentie van een deelstring stijgend in functie van de verhouding $\frac{|A|}{u}$. Dit is belangrijk aangezien het groter worden van $\frac{|A|}{u}$ betekent dat er minder deelstrings meerdere keren voorkomen. Deze deelstrings zijn bijgevolg meer discriminerend en dus ook relevant voor de vergelijking. Het gebruik van de vierkantswortel zorgt voor een niet-lineair

verloop, d.i. hoe hoger de verhouding $\frac{|A|}{u}$ wordt, hoe minder snel de drempelwaarde toeneemt. De constante 3/4 is experimenteel bepaald. De voorgestelde filter heeft de eigenschap dat deelstrings enkel worden weggefilterd als hun frequentie zeer hoog is. Met betrekking tot het voorbeeld van de deelstring "2000" in een publicatiedatabank betekent dit dat "2000" slechts wordt weggefilterd als zeer veel publicaties in de databank dateren uit het jaar 2000, waardoor de deelstring "2000" inderdaad zijn discriminerend karakter verliest. Dit zorgt ervoor dat er genoeg deelstrings overblijven om onderscheid te maken tussen coreferente en niet-coreferente strings. Immers, hoe meer strings worden weggefilterd, hoe moeilijker het wordt een kwantorfunctie te vinden die onderscheid maakt tussen coreferente en niet-coreferente strings. Als er geldt dat:

$$\frac{|A|}{u} > \frac{16}{9} \tag{6.102}$$

dan wordt geen enkele deelstring weggefilterd. Figuur 6.4 toon een grafische voorstelling van de relatieve drempel die de filter gebruikt.



Figuur 6.4: Frequentiefilter

6.6.3 Uitwisseling van kennis

Een derde en laatste aspect dat we willen behandelen, is het uitwisselen van kennis tussen evaluatoren. De evaluator $E^*_{\mathcal{S}}$ gebruikt twee evaluatieniveaus. Op het laagste niveau worden deelstrings vergeleken en op het hoogste niveau wordt de kennis van het laagste niveau gebruikt om strings te vergelijken. Het principe van kennisuitwisseling tussen evaluatoren gaat uit van het feit dat de mogelijkheid van coreferentie van een welbepaald koppel van deelstrings wordt beïnvloed door andere koppels van deelstrings. Laat ons dit verduidelijken met twee voorbeelden.

Voorbeeld 6.11
Beschouw de strings s="art's deli" en t="art's delicatessen" met $\mathscr{G} = \{\text{`}\ \text{'}\}$. *Onder deze omstandigheden geldt er dat:*

$$\begin{aligned} \mathscr{S}_{\mathscr{G}}(s) &= \{\text{“art's”}, \text{“deli”}\} \\ \mathscr{S}_{\mathscr{G}}(t) &= \{\text{“art's”}, \text{“delicatessen”}\}. \end{aligned}$$

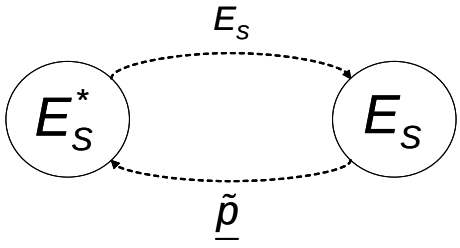
Door de sterke reflexiviteit van $E_{\mathcal{S}}$ *wordt de deelstring "art's" in s afgebeeld op de deelstring "art's" in t. Dit koppel van deelstrings evalueert naar* $(1, 0)$. *Deelstrings "deli" en "delicatessen" zijn gelijk op een suffix na. De stelling dat het suffix te wijten is aan een afkorting wordt versterkt door het feit dat een ander koppel van deelstrings aanwezig is, die met volledige zekerheid coreferent zijn.*

Voorbeeld 6.12
Beschouw de strings s="lespinasse" en t="le jardin". Onder deze omstandigheden geldt er dat:

$$\begin{aligned} \mathscr{S}_{\mathscr{G}}(s) &= \{\text{“lespinasse”}\} \\ \mathscr{S}_{\mathscr{G}}(t) &= \{\text{“le”}, \text{“jardin”}\}. \end{aligned}$$

De deelstrings "le" en "lespinasse" zijn gelijk op een suffix na, maar de stelling dat dit suffix te wijten is aan een afkorting, wordt niet bevestigd door andere koppels van deelstrings.

Dit principe zullen we toepassen op onze aanpak voor vergelijking van strings door een uitwisseling van kennis tussen de twee evaluatoren $E^*_{\mathcal{S}}$ en $E_{\mathcal{S}}$. Hierbij vertrekt de één-niveau evaluator $E_{\mathcal{S}}$ van een zeer strikt toekenningsmodel en vertrekt $E^*_{\mathcal{S}}$ van een combinatiefunctie die zeer tolerant is. Op basis van het resultaat van deze combinatie wordt een nieuw, minder streng toekenningsmodel doorgegeven aan $E_{\mathcal{S}}$, waarna de combinatiefunctie wordt verstrengd. Door steeds variërende toekenningsmodellen en combinatiefuncties wordt er iteratief een beslissing genomen. Dit principe wordt afgebeeld in Figuur 6.5. We kunnen dit proces nu meer formeel vastleggen in volgende definitie door gebruik te maken van een tweewaardig beslissingsmodel (Hoofdstuk 5).



Figuur 6.5: Uitwisseling van kennis tussen evaluatoren

Definitie 6.16 (Iteratieve berekening van onzekerheid)
Gegeven een twee-niveau evaluator $E_{\mathcal{S}}^*$ en een tweewaardig beslissingsmodel $\mathscr{B}$. Beschouw $n \in \mathbb{N}_0$ verschillende instanties van $E_{\mathcal{S}}^*$, waarbij de k^{de} instantie wordt genoteerd als $E_{\mathcal{S}}^{*,(k)}$. Veronderstel dat deze instanties voldoen aan de volgende voorwaarden. (1) Voor een willekeurige collectie van possibilistische waarheidswaarden P_π geldt er dat:

$$S_{\gamma^{T,F}}^{(k)}(P_\pi) \geq S_{\gamma^{T,F}}^{(k+1)}(P_\pi) \tag{6.103}$$

met $S_{\gamma^{T,F}}^{(k)}$ de combinatiefunctie die gebruikt wordt door $E_{\mathcal{S}}^{*,(k)}$. (2) Voor een willekeurig koppel van strings $(s,t) \in \mathcal{S}^2$ geldt er dat:

$$E_{\mathcal{S}}^{(k)}(s,t) \leq E_{\mathcal{S}}^{(k+1)}(s,t) \tag{6.104}$$

met $E_{\mathcal{S}}^{(k)}$ de gebruikte één-niveau evaluator. De onzekerheid over coreferentie kan dan iteratief worden berekend als:

$$E_{\mathcal{S}}^{*,(n)}(s,t) = \widetilde{p}^{(n)} \tag{6.105}$$

waarbij er voor willekeurige $k \in \{2,...,n\}$ geldt:

$$\widetilde{p}^{(k)} = \begin{cases} (0,1) & \textbf{als} \quad \mathscr{B}\left(\widetilde{p}^{(k-1)}\right) = F \\ E_{\mathcal{S}}^{*,(k)}(s,t) & \textbf{anders} \end{cases} \tag{6.106}$$

en waarbij:

$$\widetilde{p}^{(1)} = E_{\mathcal{S}}^{*,(1)}(s,t). \tag{6.107}$$

Een bijzonder eenvoudige maar doeltreffende instantie van dit proces is het volgende. In een eerste iteratie wordt een tweewaardige evaluator (Hoofdstuk 2) $E_{\mathcal{S}}$ voorgesteld waarmee deelstrings worden vergeleken. Dit is het strengst mogelijke toekenningsmodel voor $E_{\mathcal{S}}$ vermits enkel gelijke deelstrings coreferent kunnen zijn. $E_{\mathcal{S}}^*$ construeert op basis van deze resultaten een afbeelding en combineert de verkregen kennis met $\widetilde{\vee}$. Uit Hoofdstuk 3 weten we dat $\widetilde{\vee}$ overeenkomt met de puntsgewijs grootste instantie van $S_{\gamma^{T,F}}$. We veronderstellen een tweewaardig beslissingsmodel $\mathscr{B}$ met $z = 0$. Dit betekent dat als

het resultaat van $\widetilde{\vee}$ gelijk is aan $(0,1)$, dan is het resultaat van de iteratieve vergelijking gelijk aan $(0,1)$. Zoniet wordt het strenge toekenningsmodel voor de één-niveau evaluator voorgesteld (Tabel 6.3). De combinatiefunctie wordt dan gekozen op basis van een kwantorfunctie. Dit proces berekent de possibilistische waarheidswaarde in twee iteraties. De eerste iteratie komt neer op volgende beslissingsregel:

$$(\mathscr{S}_{\mathscr{G}}(s) \cap \mathscr{S}_{\mathscr{G}}(t) = \emptyset) \Rightarrow \left(E_{\mathcal{S}}^{*,(n)}(s,t) = (0,1)\right). \tag{6.108}$$

Het voordeel van deze strategie is dat voor heel wat niet-coreferente strings het antecedent van deze regel blijkt op te gaan. Het berekenen van de doorsnede van twee verzamelingen is echter een veel minder complexe operatie dan het evalueren van twee strings met $E_{\mathcal{S}}^{*}$. Uit onze experimenten blijkt dat deze instantie van kennisuitwisseling weinig verschil teweegbrengt in de accuraatheid van de evaluator, maar veel verschil maakt in de benodigde rekentijd (Sectie 6.8).

Voorbeeld 6.13
Beschouwen we de strings de strings s_1="art's deli", t_1="art's delicatessen", s_2="lespinasse" en t_2="le jardin". Veronderstel $\mathscr{G} = \{\text{`}\ \text{'}\}$. We vinden dan:

$$\mathscr{S}_{\mathscr{G}}(s_1) = \{\text{"art's"}, \text{"deli"}\} \tag{6.109}$$
$$\mathscr{S}_{\mathscr{G}}(t_1) = \{\text{"art's"}, \text{"delicatessen"}\} \tag{6.110}$$
$$\mathscr{S}_{\mathscr{G}}(s_2) = \{\text{"lespinasse"}\} \tag{6.111}$$
$$\mathscr{S}_{\mathscr{G}}(t_2) = \{\text{"le"}, \text{"jardin"}\}. \tag{6.112}$$

In de eerste iteratie berekenen we nu:

$$\mathscr{S}_{\mathscr{G}}(s_1) \cap \mathscr{S}_{\mathscr{G}}(t_1) \neq \emptyset \tag{6.113}$$
$$\mathscr{S}_{\mathscr{G}}(s_2) \cap \mathscr{S}_{\mathscr{G}}(t_2) = \emptyset. \tag{6.114}$$

Na de eerste iteratie vinden we dus dat:

$$E_{\mathcal{S}}^{*,(n)}(s_2,t_2) = (0,1). \tag{6.115}$$

Het is met andere woorden zeker dat s_2 en t_2 niet coreferent zijn. Voor het koppel (s_1,t_1) gebruiken we in de tweede iteratie een één-niveau evaluator met streng toekenningsmodel. Dit levert, met $c = 0.1$ (Tabel 6.3):

$$E_{\mathcal{S}}(\text{"art's"}, \text{"art's"}) = (1,0) \tag{6.116}$$
$$E_{\mathcal{S}}(\text{"deli"}, \text{"delicatessen"}) = (1,0.3). \tag{6.117}$$

Stel dat we een kwantorfunctie met parameters $(\alpha,\beta,\delta) = (1,1,1)$ kiezen, dan vinden we dat:

$$E_{\mathcal{S}}^{*,(n)}(s_1,t_1) = (1,0) \mathbin{\widetilde{\wedge}} (1,0.3) = (1,0.3). \tag{6.118}$$

6.7 Bepaling van de kwantorfunctie

In deze sectie onderzoeken we hoe de parameters van de kwantorfunctie kunnen worden bepaald. Eerst bekijken we hoe dit kan door het gebruik van een trainingscollectie. Daarna bekijken we hoe dit kan zonder het gebruik van een trainingscollectie.

6.7.1 Bepaling van een kwantorfunctie met training

Enkelvoudige kwantificatie Met de definitie van een twee-niveau evaluator voorhanden, blijft de belangrijke vraag hoe de parameters van de gebruikte kwantorfunctie vastgelegd kunnen worden, zodanig dat de evaluator een optimaal gedrag vertoont. We zullen hier eerst aandacht besteden aan een aanpak waarbij trainingsdata voorhanden is. We veronderstellen dat de kwantorfunctie die is voorgesteld in Hoofdstuk 4, wordt gebruikt. Dit wil zeggen dat drie parameters (α,β,δ) aangeleerd moeten worden door observatie van de trainingsdata. Dergelijke trainingsdata stellen we voor als een binaire en symmetrische relatie $D \subset \mathcal{S}^2$. De verzameling van alle coreferente koppels in D wordt genoteerd als C waarbij uiteraard $C \subseteq D$. We definiëren eerst de accuraatheid van een evaluator E_U als een maatstaf voor de prestatie van een evaluator.

Definitie 6.17
De accuraatheid van een evaluator E_U met betrekking tot een datacollectie $D \subseteq U^2$ wordt berekend als:

$$\theta(E_U) = \max_{\mathscr{B}} h\left(x_{\mathscr{B}}, y_{\mathscr{B}}\right) \tag{6.119}$$

waarbij:

$$x_{\mathscr{B}} = \frac{\left|\{(o_1,o_2)|(o_1,o_2) \in C \wedge \mathscr{B}\left(E_U(o_1,o_2)\right) = T\}\right|}{|C|} \tag{6.120}$$

$$y_{\mathscr{B}} = \frac{\left|\{(o_1,o_2)|(o_1,o_2) \in C \wedge \mathscr{B}\left(E_U(o_1,o_2)\right) = T\}\right|}{|\{(o_1,o_2)|(o_1,o_2) \in D \wedge \mathscr{B}\left(E_U(o_1,o_2)\right) = T\}|}. \tag{6.121}$$

Hierbij is h een monotoon stijgende functie $h : [0,1]^2 \to [0,1]$ zodat $h(1,1) = 1$.

De accuraatheid van een evaluator wordt berekend als de combinatie van twee verhoudingen. Deze beide verhoudingen staan in de literatuur beter bekend als de *compleetheid* ([99]):

$$\frac{\left|\{(o_1,o_2)|(o_1,o_2) \in C \wedge \mathscr{B}\left(E_U(o_1,o_2)\right) = T\}\right|}{|C|} \tag{6.122}$$

en de *zuiverheid* ([99]):

$$\frac{\left|\{(o_1,o_2)|(o_1,o_2) \in C \wedge \mathscr{B}\left(E_U(o_1,o_2)\right) = T\}\right|}{|\{(o_1,o_2)|(o_1,o_2) \in D \wedge \mathscr{B}\left(E_U(o_1,o_2)\right) = T\}|}. \tag{6.123}$$

Voor classificatieproblemen wordt voor de functie h vaak het harmonische gemiddelde gekozen. Het overeenkomstige criterium wordt dan de f-waarde genoemd en geeft een afweging tussen *compleetheid* en *zuiverheid*. De f-waarde wordt vaak gebruikt om de prestatie van een vergelijkingsmethode voor objecten te rapporteren en we zullen deze f-waarde dan ook gebruiken in Sectie 6.8 om onze aanpak te vergelijken met bestaande methoden. Desondanks is de f-waarde niet het criterium dat we gebruiken voor optimalisatie. We stellen hier een aanpak voorop die de *compleetheid* maximaliseert en wel om volgende redenen.

Ten eerste zullen we twee-niveau evaluatoren beschouwen in Hoofdstuk 7 als evaluatoren voor een atomair universum die deel uitmaken van een evaluator voor complexe objecten. Het kan worden ingezien dat coreferente complexe objecten enkel als dusdanig geïdentificeerd kunnen worden als ze voor de meeste van hun eigenschappen uit coreferente deelobjecten bestaan. Deze opmerking sluit aan bij de opmerking over granulariteit van de objectruimte in Hoofdstuk 2. Onder de veronderstelling van perfecte meetprocessen moet de combinatie van de kennis aangeleverd door de verschillende eigenschappen van een complex object conjunctief zijn. Wanneer meetprocessen onderhevig zijn aan imperfecties, blijft dit conjunctief karakter behouden. Dit zorgt er enerzijds voor dat een lagere *zuiverheid* van een deelevaluator gecompenseerd kan worden door de kenniscombinatie. Anderzijds impliceert een lagere *compleetheid* van een deelevaluator dat de combinatiefunctie een minder conjunctief karakter moet hebben, wat meestal leidt tot een grote foutenlast. Een twee-niveau evaluator die wordt gebruikt in een raamwerk voor detectie van coreferente complexe objecten moet bijgevolg voorkeur geven aan een hoge *compleetheid*. Optimalisatie van *zuiverheid* is van ondergeschikt belang, vermits een lagere *zuiverheid* op niveau van combinatie van kennis over eigenschappen kan worden aangepakt.

Ten tweede zal worden aangetoond dat optimalisatie van de *compleetheid* bijzonder snel en efficiënt kan, door te steunen op de eigenschappen van evaluatoren die gebruik maken van een kwantorfunctie (Hoofdstuk 4).

Ten derde zal met experimenten worden aangetoond dat optimalisatie van de *compleetheid* vaak aanleiding geeft tot hoge f-waarden in vergelijking met bestaande methoden, hetgeen wil zeggen dat de voorkeur die wordt gegeven aan *compleetheid*, vaak geen zware impact heeft op de *zuiverheid* van een evaluator. Dit betekent dat, ondanks het feit dat een twee-niveau evaluator wordt geoptimaliseerd met het oog op gebruik ervan in een evaluator voor complexe objecten, hij ook kan worden gebruikt als een alleenstaande evaluator.

Voor gegeven trainingsdata D willen we dat $E^*_{\mathcal{S}}$ zo bepaald wordt dat de *compleetheid* gelijk is aan 1. Dit betekent dat er een beslissingsmodel $\mathscr{B}$ moet bestaan zodat:

$$\forall (s,t) \in C : \mathscr{B}(E^*_{\mathcal{S}}(s,t)) = T. \tag{6.124}$$

Gelet op Definitie 5.1 is deze voorwaarde voldaan als ze voldaan is voor $\mathscr{B}$ met drempelwaarde $z = 0$. Bijgevolg moet $E^*_{\mathcal{S}}(s,t)$ een zekerheid voor waar groter dan 0 produceren voor elk koppel van strings $(s,t) \in C$. Dit betekent

dat parameter α als volgt moet worden gekozen:

$$\alpha = \max\{\alpha | \forall (s,t) \in C : \mathscr{B}(E^*_{\mathcal{S}}(s,t)) = T)\} \tag{6.125}$$

met $\mathscr{B}$ een maximaal tweewaardig beslissingsmodel. Wanneer α op deze manier wordt gekozen, kan de keuze voor β en δ nooit impliceren dat een koppel uit C een necessiteit voor waar gelijk aan 0 krijgt (zie gedeeltelijke β- en δ-invariantie in Hoofdstuk 4)[3]. Dit betekent dat β en δ onafhankelijk van α gekozen kunnen worden, zodat ze de *zuiverheid* van $E^*_{\mathcal{S}}$ maximaliseren. Experimenten leren dat de impact van δ verwaarloosbaar klein is, zodat steeds een vaste waarde 0.05 voor δ wordt vooropgesteld. Parameter β wordt gekozen zodat de *zuiverheid* maximaal wordt.

Door de specifieke eigenschappen van de kwantorfunctie kan het bepalen van de parameters (α, β, δ) een stuk efficiënter dan met een brute-kracht aanpak voor het bepalen van (α, β, δ). Dit kan worden ingezien door de volgende analyse. Drie parameters moeten worden bepaald: α, β en δ. Elk van deze parameters nemen waarden aan in het eenheidsinterval $[0, 1]$. Het aanleren gebeurt op basis van een trainingscollectie $D = A \times B$, waarbij C de verzameling van alle coreferente koppels in D voorstelt. Stel dat voor parameter $x \in \{\alpha, \beta, \delta\}$ v_x waarden getest worden[4]. Dit betekent dat een brute-kracht aanpak voor het optimaliseren van deze drie parameters een complexiteit heeft die gelijk is aan:

$$O\left(|A||B| \prod_{x \in \{\alpha,\beta,\delta\}} v_x\right) = O\left(v_\alpha v_\beta v_\delta |A||B|\right). \tag{6.126}$$

Wanneer we gebruik maken van de trainingsaanpak die steunt op de eigenschappen van de kwantorfunctie, vinden we een duidelijk lagere complexiteit. Ten eerste vereist het aanleren van parameter α enkel de coreferente koppels die typisch een zeer klein percentage van het totaal aantal koppels vormen (1-5%). Gelet op Stelling 4.3 kan α worden gevonden door binair zoeken, zodat het aantal te testen waarden van de orde $\log(v_\alpha)$ is. De complexiteit kan dan voorlopig worden herschreven als:

$$O\left(\log\left(v_\alpha\right)|C| + v_\beta v_\delta |A||B|\right). \tag{6.127}$$

Ten tweede geldt er dat koppels waarvoor $E^*_{\mathcal{S}}(s,t) < (1,1)$ onder de optimale α aan deze ongelijkheid voldoen voor alle β. Aanleren van β vereist dus alleen koppels waarvoor de evaluatie niet a priori kleiner is dan $(1,1)$. Empirische resultaten tonen aan dat het aantal koppels waarvoor de evaluatie kleiner is dan $(1,1)$ onder de optimale α typisch zeer hoog is. Dit betekent dat we de hele trainingscollectie slechts één keer moeten doorlopen voor het aanleren van

[3] Hierbij veronderstellen we impliciet dat $\delta \neq 1$.

[4] Typisch blijkt uit experimenten dat v_x van grootteorde 10 moet zijn om goede resultaten te verkrijgen.

β, zodat de factor v_β verwijderd kan worden uit de complexiteitsuitdrukking. De complexiteit kan dan herschreven worden als:

$$O\left(\log(v_\alpha)|C| + (1 + v_\delta)|A||B|\right). \tag{6.128}$$

We benadrukken dat dit duidelijk geen *slechtste-geval analyse* is, maar wel opgaat voor redelijke veronderstellingen. Ten laatste stellen we vast dat δ een vaste waarde mag toegekend worden, zonder de accuraatheid van de evaluator te benadelen. De complexiteit van het aanleren wordt dan uiteindelijk:

$$O\left(\log(v_\alpha)|C| + |A||B|\right) \approx O\left(|A||B|\right) \tag{6.129}$$

waarbij de benadering voortkomt uit $|C| << |A||B|$. De parametervector kan dus voor redelijke veronderstellingen worden aangeleerd in kwadratische tijd, hetgeen duidelijk efficiënter is dan een brute-kracht aanpak die geen rekening houdt met de eigenschappen van de kwantorfunctie.

Merk op dat de kleinst mogelijke vector van gewichten wordt bereikt als $(\alpha, \beta, \delta) = (0, 0, 0)$. In dat geval geldt dat $w_0 = 1$ en voor alle $i \neq 0$ dat $w_i \neq 1$. Dit betekent dat onze aanpak werkt als volgende veronderstelling opgaat:

$$\forall (s,t) \in C : \exists (u,v) \in \mathscr{S}_\mathscr{G}(s) \times \mathscr{S}_\mathscr{G}(t) : \mathscr{B}(E_\mathcal{S}(u,v)) = T \tag{6.130}$$

waarbij $\mathscr{B}$ een maximaal tweewaardig beslissingsmodel is (Hoofdstuk 5). Zoniet bestaan er twee coreferente strings s en t waarvoor geen enkel koppel van deelstrings enige zekerheid op coreferentie biedt. Dergelijke koppels hebben dan tot gevolg dat onze aanpak geen geschikte waarde voor α vindt. Wanneer het gemiddeld aantal deelstrings voldoende groot is, vormt deze veronderstelling meestal geen probleem. Bij een gemiddeld klein aantal deelstrings, kan de veronderstelling in de praktijk niet opgaan. Daarvoor bestaan twee belangrijke redenen.

Ten eerste kan het zijn dat de trainingscollectie fouten bevat, d.i. twee strings zijn coreferent maar dit kan voor geen enkele keuze van de parameters worden vastgesteld door de evaluator. Een dergelijke fout kan optreden omdat data niet up-to-date zijn, bijvoorbeeld adresgegevens van een persoon die niet aangepast zijn. Ook kan het zijn dat de evaluator niet geschikt is om te oordelen over coreferentie, bijvoorbeeld omdat eerder een semantische evaluator (Hoofdstuk 2) nodig is. Wanneer dit het geval is kan de bepaling van α door (6.125) worden aangepast zodat een hoog percentage van koppels uit C een necessiteit voor waar groter dan 0 toegekend krijgen, eerder dan alle koppels uit C. In het algemeen kan een dergelijke strategie de robuustheid van de verkregen evaluator ten goede komen, maar dan vervalt het uitgangspunt van maximale *compleetheid*.

Ten tweede kan de splitsingsfunctie aan de basis van het probleem liggen. Beschouw het voorbeeld $s =$ "groteweg" en $t =$ "grote weg" en een tweewaardige één-niveau evaluator $E_\mathcal{S}$, dan kan een evaluator $E^*_\mathcal{S}$ geen zekerheid voor waar vinden, ongeacht de waarde voor parameter α. Dit probleem kan worden aangepakt door meer geavanceerde splitsingsfuncties te beschouwen of door deelstrings stelselmatig te concateneren (Sectie 6.6).

Meervoudige kwantificatie Aan het einde van Hoofdstuk 4 is aangehaald dat het gebruik van een eenvoudige kwantorfunctie niet altijd volstaat om het onderscheid tussen coreferente en niet-coreferente collecties te specificeren. Om dit probleem aan te pakken is voorgesteld om de parameters (α, β, δ) afhankelijk te maken van de te vergelijken collecties. Deze oplossing noemen we meervoudige kwantificatie. Willen we het nut van deze aanpak hier nagaan, dan moeten we het aanleren van een parametervector uitbreiden naar de context van meervoudige kwantificatie. We veronderstellen m klassen in het eenheidsinterval (Hoofdstuk 4), waardoor de parameters kunnen worden voorgesteld in een parametermatrix $\mathbf{M}_q$:

$$\mathbf{M}_q = \begin{bmatrix} b_1 & \alpha_1 & \beta_1 & \delta_1 \\ b_2 & \alpha_2 & \beta_2 & \delta_2 \\ & & \ldots & \\ b_m & \alpha_m & \beta_m & \delta_m \end{bmatrix}. \tag{6.131}$$

De eerste kolom bevat de bovengrenzen b_i van de klassen K_i en dus geldt er dat $b_m = 1$. Het belangrijkste verschil met het beschreven leerproces van een parametervector is dat de bovengrenzen b_i en dus ook de klassen K_i eerst bepaald moeten worden. Bij de aanpak die we hier volgen, veronderstellen we dat de eerste klasse ($[0, b_1]$) geassocieerd wordt met een vaste parametervector zodat $\alpha_1 = 1$, $\beta_1 = 1$ en $\delta_1 = 0.05$. Dit betekent dat koppels waarvoor de kardinaliteitsverhouding kleiner is dan b_1 een maximum aan coreferente deelstrings moeten bezitten opdat ze als coreferent bestempeld zouden worden. Bijgevolg kan b_1 worden berekend als:

$$b_1 = \max\left\{b | \forall (s,t) \in C : \mathrm{kv}_{s,t} \leq b \Rightarrow \mathscr{B}(E^*_{\mathcal{S},1}(s,t)) = T\right\} \tag{6.132}$$

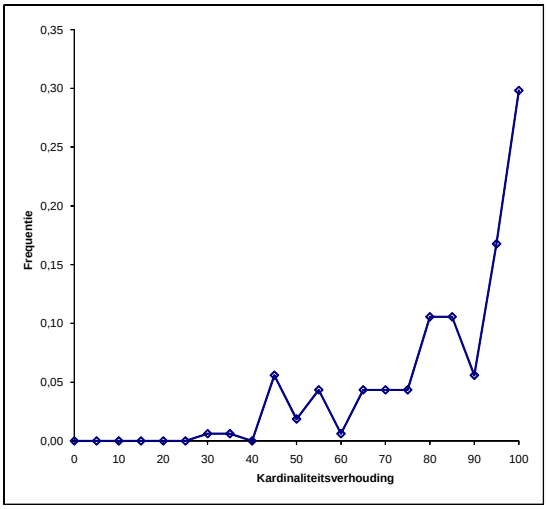
waarbij $\mathscr{B}$ een maximaal beslissingsmodel is en waarbij $\mathrm{kv}_{s,t}$ de kardinaliteitsverhouding van s en t voorstelt:

$$\mathrm{kv}_{s,t} = \frac{\min(|\mathscr{S}_{\mathscr{G}}(s)|, |\mathscr{S}_{\mathscr{G}}(t)|)}{\max(|\mathscr{S}_{\mathscr{G}}(s)|, |\mathscr{S}_{\mathscr{G}}(t)|)}. \tag{6.133}$$

De bepaling van $b_2, ..., b_{m-1}$ kan op verschillende manieren gebeuren. Echter, de keuze van de klassen is belangrijk voor het vermijden van *overfitting*. Onze aanpak kiest klassen zodat het aantal geobserveerde coreferente koppels quasi-uniform wordt verdeeld over de verschillende klassen. Op die manier kan elke kwantorfunctie worden aangeleerd op een gelijk aantal voorbeelden van coreferente koppels.

Figuur 6.6 toont de verdeling van kardinaliteitsverhoudingen van coreferente koppels voor één van de datacollecties uit Sectie 6.8. Hieruit blijkt dat coreferente koppels van strings typisch een hoge kardinaliteitsverhouding hebben. Deze typische asymmetrie heeft een belangrijk gevolg. Het betekent namelijk dat voor $i < j$, K_i typisch een grotere breedte zal hebben dan K_j, waardoor het aantal mogelijke klassen m een praktische bovengrens kent.

Parameters α_i voor $i \in \{2, ..., m\}$ kunnen worden aangeleerd op de voorbeelden die een kardinaliteitsverhouding in klasse K_i hebben. Dit gebeurt op



Figuur 6.6: Verdeling van de kardinaliteitsverhouding

dezelfde manier als eerder beschreven voor enkelvoudige kwantificatie. Elke α_i kan onafhankelijk worden aangeleerd vermits enkel de coreferente koppels nodig zijn. Dit geldt echter niet voor het aanleren van parameters β_i. Het is namelijk zo dat de optimalisatie van β_i op klasse K_i geen globaal optimale β_i impliceert. Dit komt omdat elke β_i zo berekend moet worden, dat een hogere possibiliteit wordt toegekend aan coreferente koppels. Dit betekent echter dat alle niet-coreferente koppels in rekening gebracht moeten worden. Hierbij aansluitend kan worden opgemerkt dat optimalisatie van de verschillende β_i's géén convex probleem is. Deze argumenten impliceren dat het aantal klassen m niet te hoog mag worden gekozen als de complexiteit van de trainingsfase belangrijk is.

6.7.2 Bepaling van een kwantor zonder training

Enkelvoudige kwantificatie We beschikken nu over een methode om een kwantorfunctie te trainen op basis van een trainingscollectie D. In praktische toepassingen is het opbouwen van een dergelijke collectie echter een dure operatie. Dit is te wijten aan het feit dat enerzijds typisch $|C| << |D|$ en anderzijds net C belangrijk is voor het trainen[5]. Het opbouwen van de trainingscollectie vereist bijgevolg dat een deel van het coreferentieprobleem handmatig wordt opgelost. Daarom is het vaak nuttig of zelfs noodzakelijk een methode te gebruiken die de parameters van de kwantorfunctie bepaalt zonder het gebruik van een trainingscollectie.

Onze aanpak voor de bepaling van parameters vertrekt van de vaststel-

[5]Dit geldt voor zowel de aanpak die we hier hebben beschreven als voor andere methoden uit de literatuur.

ling dat voor een gegeven datacollectie $D \subseteq \mathcal{S}^2$ typisch geldt dat $|C| << |D|$. Aangezien enkel parameter α bepaalt welke waarheidswaarde een necessiteit verschillend van nul toegewezen krijgt, zal onze parameterbepaling zich richten op deze parameter. Voor de andere parameters zal een heuristiek worden gebruikt. We herhalen dat $E^*_{\mathcal{S}}$ stijgend is voor dalende α. Elke mogelijke waarde $v \in [0,1]$ voor α geeft aanleiding tot een kandidaatverzameling $\widehat{C}_v$ als volgt:

$$\widehat{C}_v = \{(s,t)|(s,t) \in D \wedge \mathscr{B}(E_{\mathcal{S},\alpha=v}(s,t)) = T\}. \tag{6.134}$$

Hierbij is $\mathscr{B}$ een maximaal tweewaardig beslissingsmodel. Dit betekent dat $|\widehat{C}_v|$ gelijk is aan het aantal koppels waarvoor een zekerheid voor coreferentie verschillend van 0 bestaat. Gelet op de definitie van een tweewaardig beslissingsmodel (Definitie 5.1), kan $|\widehat{C}_v|$ worden gezien als een bovengrens voor $|\widehat{C}|$. Er geldt voor elke twee waarden $v \in [0,1]$ en $v' \in [0,1]$ dat:

$$v \geq v' \Leftrightarrow \widehat{C}_v \subseteq \widehat{C}_{v'}. \tag{6.135}$$

We maken nu de veronderstelling dat de verhouding:

$$\frac{\left|C \cap \widehat{C}_v\right|}{|\widehat{C}_v|} \tag{6.136}$$

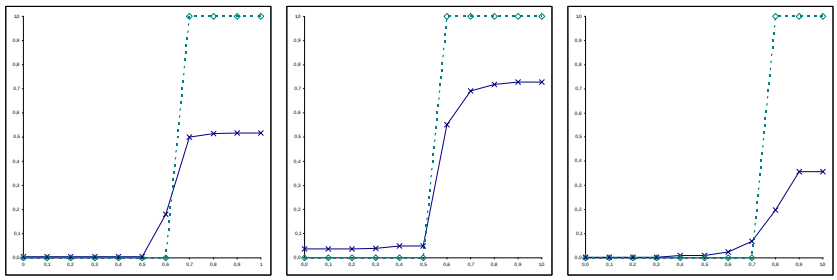
dalend is voor dalende v. Deze veronderstelling houdt in dat, naarmate er meer koppels worden gevonden met necessiteit voor coreferentie verschillend van 0, steeds meer van deze koppels niet coreferent zijn. Anders gezegd, de *zuiverheid* van de verzameling $\widehat{C}_v$ met betrekking tot coreferente objecten [99] is dalend in functie van dalende v. Hierop gelet stellen we ons de vraag of de *zuiverheid* kan worden bepaald. Een absolute bepaling is moeilijk aangezien $|C|$ onbekend is en het moeilijk is een schatter te construeren die algemeen geldig is. Het is echter wel mogelijk een relatieve bepaling te maken, d.i. een bepaling van het verloop van de *zuiverheid*. Eerder dan een bepaling te maken van de eigenlijke *zuiverheid*, zullen we nagaan of de *zuiverheid* nog voldoende hoog is. We stellen vast dat de daling in *zuiverheid* voor $|v - v'|$ kan worden benaderd door $||\widehat{C}_v| - |\widehat{C}_{v'}||$. De volgende methode wordt voorgesteld om het verloop van de *zuiverheid* te bepalen. We kiezen $v_0 = 1$, we berekenen $\widehat{C}_{v_0}$ en we veronderstellen dat $e_0 = 1$. Voor elke v_i berekenen we:

$$e_i = \begin{cases} 1 & \textbf{als} \quad |\widehat{C}_{v_{i-1}}| = 0 \wedge e_{i-1} = 1 \\ 1 & \textbf{als} \quad \frac{|\widehat{C}_{v_i}|}{|\widehat{C}_{v_{i-1}}|} < \tau \wedge e_{i-1} = 1 \\ 0 & \textbf{anders} \end{cases} \tag{6.137}$$

waarbij τ een drempelwaarde is die het proces stuurt. Voor deze drempelwaarde kan een constante waarde worden gevonden onafhankelijk van de datacollectie. Tijdens experimenten wordt $\tau = 2$ gekozen. De bepaling van parameter α gebeurt dan als volgt:

$$\alpha = \min\{v_i | v_i \in [0,1] \wedge e_i = 1\}. \tag{6.138}$$

De waarde voor α is dus de kleinste v_i waarvoor $e_i = 1$. Dit wil zeggen dat α zo klein mogelijk wordt gekozen, zodat de *zuiverheid* nog voldoende groot is. Dit principe wordt geïllustreerd in Figuur 6.7 voor drie datacollecties, waar de onderbroken lijnen telkens het verloop van de e_i's voorstelt. De volle lijn toont de eigenlijke *zuiverheid*.



Figuur 6.7: Bepaling van α: *zuiverheid* (volle lijn) en gereconstrueerde *zuiverheid* (onderbroken lijn)

De waarden voor parameters β en δ worden vast gekozen in de aanpak zonder trainingscollectie. Voor δ is dit 0.05, naar analogie met de aanpak op basis van een trainingscollectie. Voor β is dit typisch 0.

Meervoudige kwantificatie Gelet op de invariabiliteit van het merendeel van de parameters in het geval van twee kwantorfuncties, is het ook mogelijk om twee kwantorfuncties te construeren zonder trainingscollectie. De eerste kwantorfunctie wordt hierbij vast gekozen en bijgevolg moet bij deze enkel de splitsingswaarde van de intervallen (d.i. b_1) worden bepaald. Uit experimenten rond het bepalen van twee kwantorfuncties op basis van een trainingscollectie wordt vastgesteld dat de waarde b_1 bijna invariabel is met betrekking tot verschillende datacollecties. Dit betekent dat we ook b_1 een vaste waarde kunnen toekennen. Voor experimenten rond de accuraatheid van evaluatoren gebaseerd op deze parameterbepaling, verwijzen we naar Sectie 6.8.

6.8 Experimenten

In deze sectie wordt de accuraatheid van onze aanpak onderzocht aan de hand van vergelijkende experimenten met bestaande methoden. We gaan uit van een datacollectie D die bestaat uit koppels van strings. Er geldt dus dat $D \subset \mathcal{S}^2$. We noteren de verzameling van alle coreferente koppels in D als C. Een benadering van de verzameling van coreferente koppels noteren we als $\widehat{C}$, hetgeen ons toelaat twee maten te definiëren, namelijk *zuiverheid*:

$$\text{zuiv} = \frac{|C \cap \widehat{C}|}{|\widehat{C}|} \tag{6.139}$$

en *compleetheid*:

$$\text{comp} = \frac{|C \cap \widehat{C}|}{|C|}. \tag{6.140}$$

De accuraatheid van een evaluator wordt uitgedrukt door middel van de f-waarde, die het harmonische gemiddelde is van de *zuiverheid* en de *compleetheid*:

$$f = \frac{2 \text{ zuiv comp}}{\text{zuiv} + \text{comp}} \tag{6.141}$$

Voor de rapportering van onze resultaten stellen we tien waarden voor *compleetheid* voorop $(0.1, ..., 1.0)$ en berekenen we voor elk van deze waarden de grootst mogelijke *zuiverheid*. Dit doen we door koppels (s, t) te rangschikken volgens grootste $E_{\mathcal{S}}(s, t)$ (of $E^*_{\mathcal{S}}(s, t)$) eerst. Voor een vooropgestelde waarde van *compleetheid* kan deze rangschikking worden overlopen tot het vereiste aantal coreferente koppels wordt geobserveerd. Koppels met dezelfde evaluatie onder $E_{\mathcal{S}}$ worden samen geobserveerd vermits $E_{\mathcal{S}}$ geen onderscheid kan maken tussen deze koppels. Wanneer het vereiste aantal coreferente koppels wordt geobserveerd, kan de *zuiverheid* worden berekend en deze *zuiverheid* is de hoogst mogelijke voor de vooropgestelde *compleetheid*. We zullen de maximale f-waarde berekenen over alle vooropgestelde waarden van *compleetheid*. Wanneer nodig zal de *zuiverheid* tegenover de *compleetheid* worden uitgetekend. We merken op dat deze methoden voor rapportering de standaardmethoden uit de literatuur zijn. De datacollecties die worden gebruikt tijdens onze experimenten zijn opgesomd in Tabel 6.5. Deze tabel toont de naam van de datacollectie, de wetenschappelijke bron waar de datacollectie voor het eerst is vermeld, het aantal koppels, het aantal coreferente koppels en het gemiddeld aantal deelstrings na splitsing. Datacollectie 'census' komt voort uit de persoonlijke communicatie van Winkler met zijn collega's en datacollectie 'streets5' is een eigen ontworpen datacollectie. Voor enkele steekproeven uit deze datacollecties verwijzen we naar Bijlage E.

Om een eerlijke vergelijking te bekomen zullen we onze één-niveau aanpak vergelijken met enkele karaktergebaseerde methoden uit de literatuur. Daarna worden twee-niveau methoden onderling vergeleken. Hierbij wordt de invloed onderzocht van de technieken geïntroduceerd in dit hoofdstuk.

6.8.1 Eén-niveau evaluatie

In een eerste reeks van experimenten wordt het gedrag van de één-niveau evaluator bestudeerd. We zullen onze aanpak vergelijken met een aantal methoden uit de literatuur die zorgvuldig geselecteerd zijn. Eerst en vooral vermelden we de resultaten gebaseerd op de Levenshtein afstand, omwille van zijn fundamentele rol in de literatuur over stringvergelijking. Daarnaast vermelden we de resultaten van de karaktergebaseerde similariteitsmaat van Jaro, omwille van zijn robuustheid zoals aangegeven door [88]. De Jaro-score van twee strings s en t wordt als volgt berekend. Eerst wordt gezocht naar de gemeenschappelijke karakters van de strings. Het karakter $\tau_s(i)$ is gemeenschappelijk met

naam	**bron**	$\lvert D\rvert$	$\lvert C\rvert$	gem. $\lvert\mathscr{S}_{\mathscr{G}}(s)\rvert$
people	[85]	2025	45	3
bird1	[100]	6340	19	4
bird2	[100]	62152	66	2
bird3	[100]	345	15	2
bird4	[100]	87420	155	4
game	[100]	86982	45	4
park	[100]	102168	250	3
animal	[100]	4671810	229	2
census	Winkler	176008	327	4
univ	[85]	6670	676	5
streets5	-	852930	350	1
constraint	[101]	43365	161	24
face	[101]	60726	294	25
reasoning	[101]	131841	756	24
reinforcement	[101]	82215	827	25
restaurant	[102]	176423	112	2

Tabel 6.5: Datacollecties voor $\mathcal{S}$

t als er een karakter $\tau_t(j) = \tau_s(i)$ bestaat in t zodat $i - H \leq j \leq i + H$ en $H = \min(|s|, |t|)/2$. Laat s' de string van karakters zijn die s gemeenschappelijk heeft met t en t' de string van karakters die t gemeenschappelijk heeft met s. In de zin van Jaro is een transpositie tussen s' en t' een positie i zodat $\tau_{s'}(i) \neq \tau'_t(i)$. Indien $X_{s',t'}$ de helft van het aantal transposities tussen s' en t' voorstelt, dan is de Jaro-score gegeven als:

$$\text{Jaro}(s,t) = \frac{1}{3}\left(\frac{|s'|}{|s|} + \frac{|t'|}{|t|} + \frac{|s'| - X_{s',t'}}{|s|}\right). \tag{6.142}$$

Deze berekening heeft tot gevolg dat elk karakter van s moet worden opgezocht in een deelstring van t met lengte H en omgekeerd voor elk karakter van t. Hieruit volgt dat de aanpak van Jaro kwadratisch is in het slechtste geval. Hoewel de meeste karaktergebaseerde methoden een kwadratische complexiteit hebben, kan de aanpak van Jaro veel efficiënter worden geïmplementeerd. Ten slotte wordt de methode van Monge en Elkan gerapporteerd aangezien deze methode als de beste karaktergebaseerde aanpak wordt gezien door [88]. De methode van Monge en Elkan is een variant van de editeerafstand in die zin dat operaties op blokken van karakters tegen een lagere kost kunnen gebeuren dan de som van de operaties op de individuele karakters. Een dergelijke aanpak wordt ook een *affiene-gaten afstand* genoemd. Bovendien beschouwen Monge en Elkan benaderende overeenkomsten tussen de volgende groepen van karakters: $\{d, t\}$, $\{g, j\}$, $\{l, r\}$, $\{m, n\}$, $\{b, p, v\}$ en $\{a, e, i, o, u\}$. Dit implementeren ze door de kost tussen deze karakters te verlagen tijdens de berekening van de editeerafstand. Voor meer details verwijzen we naar [85].

Voor de karaktergebaseerde methoden zullen we ons op het vlak van ac-

naam	Levenshtein	Jaro	Monge-Elkan	$E_{\mathcal{S}}$
people	0.2069	0.2507	**0.7417**	0.0435
bird1	0.5893	0.8235	**0.9474**	0.0060
bird2	0.8815	0.8815	**0.9179**	0.8815
bird3	0.3000	0.2727	**0.8103**	0.0833
bird4	0.9016	0.4571	**0.9410**	0.0035
game	0.7500	0.7297	**0.7554**	0.7200
park	0.9260	0.9337	**0.9395**	0.9375
animal	0.0142	**0.1204**	0.0068	0.0001
census	**0.8093**	0.6824	0.7004	0.4594
univ	0.6865	**0.7582**	0.6437	0.4615
streets5	0.9721	0.9474	0.8221	**0.9790**
constraint	0.7782	0.7327	**0.8508**	0.1818
face	0.8235	0.7106	**0.8367**	0.1818
reasoning	**0.7679**	0.6434	0.7022	0.0114
reinforcement	**0.7980**	0.7047	0.7834	0.1818
restaurant	**0.8192**	**0.8192**	0.7884	0.8192

Tabel 6.6: Maximale f-waarden: karaktergebaseerde methoden

curaatheid beperken tot het vergelijken van de maximale f-waarden. Deze waarden zijn voldoende indicatief voor de verschillen tussen datacollecties en methoden. Tabel 6.6 toont de maximale f-waarden van de één-niveau evaluator vergeleken met de drie methoden uit de literatuur. Voor elke datacollectie wordt de beste aanpak in vet gezet. Het valt onmiddellijk op dat de aanpak van Monge en Elkan superieur is aan de andere methoden, hetgeen niet echt mag verbazen gezien hun verfijnd model voor vergelijking van strings dat rekening houdt met afkortingen en schrijffouten. Ook merken we op dat de één-niveau aanpak gemiddeld gezien niet goed presteert in termen van maximale f-waarde. Deze resultaten dienen echter vanuit het juiste standpunt te worden geïnterpreteerd. De twee karakteristieke kenmerken van onze één-niveau aanpak zijn immers een bijzonder strikte toekenning van possibiliteiten en een lage complexiteit. De strikte toekenning impliceert dat een datacollectie met strings enkel kan worden beoordeeld door een één-niveau evaluator als het foutenmodel van de evaluator van toepassing is op de datacollectie. Onze eigen datacollectie 'streets5' is bijvoorbeeld ontwikkeld door een willekeurige foutenintroductie voorop te stellen, waarbij de fouten overeenkomen met de editeeroperaties invoeging, verwijdering en vervanging. We zien duidelijk dat onder deze (strikte) voorwaarden onze aanpak goed presteert. Heel wat datacollecties vallen echter buiten het strikte foutenmodel dat wordt verondersteld door de één-niveau evaluator, zodat het niet hoeft te verbazen dat de resultaten voor deze collecties slecht zijn. De idee hierachter is dat een één-niveau evaluator typisch als een component van een twee-niveau evaluator werkt. Zoals uit volgende resultaten zal blijken, volstaat het strikte foutenmodel van de één-niveau evaluator meestal om deelstrings te kunnen vergelijken. Een bijkomend voordeel van

onze aanpak is de bijzonder lage complexiteit. Tabel 6.7 toont de som van de uitvoeringstijden overheen alle datacollecties voor de vier methoden. Het verschil in uitvoeringstijd voor de datacollecties is enorm. Bovendien is voor alle datacollecties behalve 'people', 'bird1' en 'bird3' afzonderlijk vastgesteld dat de één-niveau evaluator een significant snellere uitvoeringstijd vertoont dan alle andere methoden. Om dit te onderzoeken zijn de berekeningen herhaaldelijk uitgevoerd en is een t-toets ($p < 0.05$) gebruikt om gemiddelde uitvoeringstijden te vergelijken. Dat er voor de datacollecties 'people', 'bird1' en 'bird3' geen significante verschillen worden vastgesteld, is te wijten aan het lage aantal te vergelijken koppels (Tabel 6.5, kolom 3).

aanpak	**uitvoeringstijd (sec)**
Levenshtein	1190
Jaro	102
Monge-Elkan	6090
$E_{\mathcal{S}}$	82

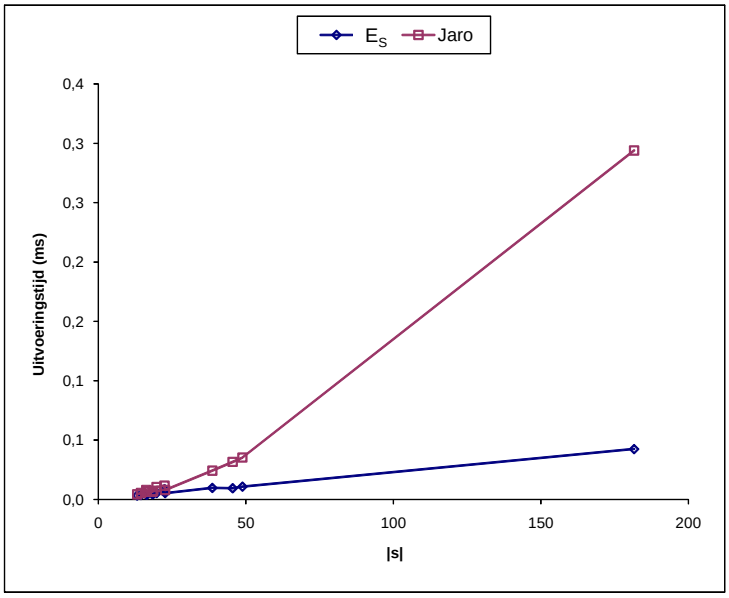
Tabel 6.7: Som van de uitvoeringstijden voor alle datacollecties

Om de complexiteit van onze één-niveau aanpak nog beter te bestuderen, is de gemiddelde uitvoeringstijd berekend in functie van de gemiddelde stringlengte. Het resultaat hiervan wordt vergeleken met de aanpak van Jaro in Figuur 6.8. Hoewel voor beide methoden is geweten dat ze een kwadratische complexiteit hebben in het slechtste geval, toont Figuur 6.8 een duidelijk groter wordend verschil in uitvoeringstijd naarmate de stringlengte stijgt. De conclusie van de experimenten met de één-niveau aanpak kan worden samengevat door te stellen dat onze één-niveau aanpak een lage complexiteit vertoont, maar door de bijzonder strenge toekenning van possibiliteit op coreferentie aan koppels van strings is de techniek enkel toepasbaar in een beperkt kader.

6.8.2 Twee-niveau evaluatie

De voordelen van de één-niveau aanpak komen ten volle tot hun recht in de context van de voorgestelde twee-niveau aanpak. Uit de literatuur van twee-niveau systemen kiezen we de SoftTFIDF aanpak, die algemeen bekend staat als de beste aanpak in termen van maximale f-waarde [88]. Laat ons deze techniek wat nader toelichten. SoftTFIDF is een twee-niveau aanpak die steunt op een tweede vergelijkingstechniek voor strings, die we hier noteren als sim. In de literatuur wordt typisch sim =Jaro-Winkler gekozen [88]. In wat volgt maken we ook hier deze keuze. Er wordt een splitsingsfunctie $\mathscr{S}_{\mathscr{g}}$ beschouwd en op basis van de resulterende multiverzamelingen wordt de volgende verzameling geconstrueerd:

$$Y(\theta, s, t) = \{r | r \in \mathscr{S}_{\mathscr{g}}(s) \wedge \exists r' \in \mathscr{S}_{\mathscr{g}}(t) : sim(r, r') > \theta\}\,. \tag{6.143}$$



Figuur 6.8: Gemiddelde uitvoeringstijd in functie van de stringlengte

Typisch wordt $\theta = 0.9$ gekozen [88], een keuze die ook hier wordt genomen. Vervolgens definieert men:

$$\forall r \in Y(\theta, s, t) : m(r) = \arg \max_{r' \in \mathscr{S}_{\mathscr{G}}(t)} sim(r, r'). \tag{6.144}$$

Met deze notaties kan SoftTFIDF(s, t) worden berekend als volgt:

$$\sum_{r \in Y(\theta,s,t)} \frac{\mathbf{v}_r^s \mathbf{v}_{m(r)}^t}{\sqrt{\left(\sum_{r \in \mathscr{S}_{\mathscr{G}}(s)} \left(\mathbf{v}_r^s\right)^2\right)} \sqrt{\left(\sum_{r \in \mathscr{S}_{\mathscr{G}}(t)} \left(\mathbf{v}_r^t\right)^2\right)}} sim(r, m(r)). \tag{6.145}$$

waarbij $\mathbf{v}^s$ en $\mathbf{v}^t$ gewichtsvectoren voor s en t zijn, geconstrueerd volgens het TFIDF principe (zie de inleiding van dit hoofdstuk). SoftTFIDF is dus gelijk aan de cosinus similariteit van deze vectoren, lokaal gewogen met de similariteit tussen twee vectordimensies.

Basistest In een eerste basistest vergelijken we de SoftTFIDF techniek met onze aanpak, door twee naïeve keuzes voor de kwantorfunctie te maken. Meer bepaald stellen we twee keer een vaste kwantorfunctie voorop waarbij $\alpha = 1$ en $\delta = 0.05$. De twee kwantorfuncties verschillen enkel in de waarde die ze voor parameter β aannemen: voor de eerste kwantorfunctie is $\beta = 0$ en voor de tweede is $\beta = 1$. De gebruikte één-niveau evaluator is gebaseerd op het strenge toekenningsmodel. Verder veronderstellen we het gebruik van een

frequentiefilter en de uitwisseling van kennis tussen de evaluatoren, d.i. door de filter uit Sectie 6.6 toe te passen. De maximale f-waarden voor SoftTFIDF worden vergeleken met de maximale f-waarden voor deze kwantorfuncties in Tabel 6.8. In elke rij wordt de beste methode in het vet gezet.

naam	**SoftTFIDF**	$E^*_{\mathcal{S}}(\beta = 0)$	$E^*_{\mathcal{S}}(\beta = 1)$
people	0.9474	**1.0000**	0.9890
bird1	0.9217	**0.9474**	**0.9474**
bird2	0.8970	0.9179	**0.9251**
bird3	0.8889	**0.9474**	**0.9474**
bird4	0.9316	**0.9474**	**0.9474**
game	**0.8276**	0.6802	0.7214
park	**0.9356**	0.8811	0.8678
animal	0.7452	0.8639	**0.9138**
census	**0.6857**	0.1789	0.1789
univ	**0.8615**	0.6335	0.6635
streets5	**0.9474**	**0.9474**	**0.9474**
constraint	**0.9116**	0.5714	0.5714
face	**0.9022**	0.5646	0.5687
reasoning	**0.9009**	0.6614	0.6632
reinforcement	**0.8780**	0.4615	0.4615
restaurant	0.8392	0.8192	**0.8624**

Tabel 6.8: Maximale f-waarden: vaste kwantorfunctie

We herhalen dat $\alpha = 1$ betekent dat twee strings s en t als coreferent gezien worden als alle deelstrings van de ene string coreferent zijn met een deelstring uit de andere. Voor een grafische voorstelling van de kwantorfunctie in het geval $\alpha = 1$ en $\beta = 0$ verwijzen we naar Figuur 4.8. We herinneren ook aan de gedeeltelijke β-invariantie besproken in Hoofdstuk 4. Hierop gelet leiden we uit Tabel 6.8 af dat een naïeve keuze voor de parameters van de kwantorfunctie leidt tot relatief accurate resultaten. Er blijkt echter duidelijk dat de naïeve aanpak faalt voor datacollecties met een hoog gemiddeld aantal deelstrings na splitsing. De resultaten tonen echter aan dat het coreferent zijn van strings in heel wat situaties ontdekt kan worden door een algemeen model. Anders gezegd, hoewel de conditionele necessiteit voor twee strings s en t enkel afhangt van $|\mathscr{S}_{\mathscr{G}}(s)|$ en $|\mathscr{S}_{\mathscr{G}}(t)|$ en niet van de eigenlijke deelstrings van s en t, laat deze conditionele necessiteit toe om onzekerheid over coreferentie te gaan bepalen. Hiermee raken we aan het belangrijke verschil tussen SoftTFIDF en onze possibilistische aanpak. Een methode zoals SoftTFIDF zal voor elke deelstring een gewicht bepalen, waardoor het vergelijken van twee strings s en t gebeurt op basis van een aangepaste gewichtsvector. De possibilistische aanpak daarentegen gebruikt eenzelfde kwantorfunctie om twee willekeurige strings te vergelijken. Het gebruik van een dergelijk algemeen model is tot op heden nergens beschreven in de literatuur, maar de eerste resultaten laten vermoeden dat de achterliggende idee werkt. We zullen overigens aantonen dat voor data-

collecties waarvoor de naïeve aanpak faalt, een aangepaste kwantorfunctie kan worden bepaald, met betere resultaten tot gevolg.

Parameterbepaling Er wordt nu bestudeerd wat het gevolg is van het trainen van een kwantorfunctie op basis van een trainingscollectie en het bepalen van een kwantorfunctie zonder trainingscollectie. Ook houden we nu rekening met het feit dat de splitsing van een string in deelstrings soms onderhevig is aan fouten, zoals aangehaald in Sectie 6.6. Dit impliceert dat bij het vergelijken van twee strings s en t, deelstrings van s worden samengevoegd als hun concatenatie voorkomt in $\mathscr{S}_{\mathscr{q}}(t)$ en omgekeerd. We brengen deze stap hier in rekening omdat is aangehaald dat fouten van de splitsingsfunctie een negatief effect kunnen hebben op de parameterbepaling. Om het bepalen van parameters aan de hand van een trainingscollectie mogelijk te maken, worden de datacollecties 50 maal opgesplitst in twee gelijke delen. Het eerste deel wordt gebruikt om parameters te bepalen, het tweede deel wordt gebruikt voor de evaluatie. De gemiddelde maximale f-waarde wordt vergeleken met de maximale f-waarde verkregen (1) door toepassing van SoftTFIDF en (2) door toepassing van een evaluator met parameters bepaald zonder trainingscollectie. De beste methode wordt in het vet gezet en een *-symbool duidt op een significant verschil met de gemiddelde maximale f-waarde bij training. Deze significanties zijn het resultaat van een t-toets. Voor datacollectie 'people' kon geen t-toets worden berekend aangezien de standaardafwijking van de reeks f-waarden gelijk is aan 0.

naam	**SoftTFIDF**	$E^*_{\mathcal{S}}$ (training)	$E^*_{\mathcal{S}}$ (benadering)
people	0.9474	**1.0000**	0.9890
bird1	0.9217	**0.9227**	0.9217
bird2	0.8970*	**0.9100**	0.8815*
bird3	0.8889*	0.9396	**0.9474**
bird4	0.9316	**0.9418**	0.9254*
game	**0.8276***	0.8063	0.7381*
park	**0.9356***	0.8443	0.9093*
animal	0.7452*	0.8286	**0.8787***
census	**0.6857***	0.5231	0.1789*
univ	**0.8615***	0.8249	0.8281
streets5	0.9474	**0.9482**	0.9474
constraint	0.9116	0.9210	**0.9262**
face	0.9022*	0.9321	**0.9340**
reasoning	**0.9009**	0.9001	0.8811
reinforcement	0.8780*	**0.9053**	0.4615
restaurant	0.8392*	**0.8583**	0.8192

Tabel 6.9: Maximale f-waarden: aangepaste kwantorfunctie

De resultaten van het experiment met een aangepaste kwantorfunctie tonen aan dat de possibilistische aanpak een beter resultaat geeft voor een aantal datacollecties. In het bijzonder blijkt dat een aangepaste kwantorfunctie een

naam	% koppels	% coreferente koppels
people	94.96	0.00
bird1	93.80	0.00
bird2	96.38	0.00
bird3	94.20	6.67
bird4	96.37	0.00
game	96.80	2.22
parks	86.73	0.40
animal	99.49	0.87
census	90.52	0.00
univ	74.06	8.57
streets5	99.46	4.01
constraint	6.31	0.00
face	10.00	0.00
reasoning	0.18	0.00
reinforcement	0.12	0.00
restaurant	98.30	0.00

Tabel 6.10: Percentage van gefilterde (coreferente) koppels bij uitwisseling van kennis

sterk effect heeft voor datacollecties met lange strings. Toch stellen we vast dat voor vier datacollecties SoftTFIDF een significant beter resultaat geeft. We zullen dit in volgende experimenten verder onderzoeken. Daarbij zullen we de invloed van een aantal besproken technieken nader onderzoeken. Hierbij zullen we ons voornamelijk toeleggen op de parameterbepaling zonder trainingscollectie vermits deze aanpak in Hoofdstuk 7 zal worden gebruikt in het geval van complexe objecten.

De invloed van kennisuitwisseling Zoals uitgelegd hebben we tot hiertoe een uitwisseling van kennis tussen evaluatoren verondersteld. Dit betekent dat in een eerste iteratie een filter wordt toegepast die voor twee strings s en t de volgende regel verifieert:

$$\mathscr{S}_{\mathscr{G}}(s) \cap \mathscr{S}_{\mathscr{G}}(t) = \emptyset \Rightarrow E^*_{\mathcal{S}}(s,t) = (0,1). \tag{6.146}$$

Het voordeel hiervan is dat de berekening van $E^*_{\mathcal{S}}(s,t)$ voor heel wat koppels kan worden vermeden. Gelet op de kwadratische complexiteit van het vergelijken van (multi)verzamelingen kan dit een significante winst in rekentijd betekenen. Tabel 6.10 toont het percentage van koppels dat wordt gefilterd en vergelijkt dit met het percentage coreferente koppels dat wordt gefilterd. Deze resultaten tonen aan dat voor de datacollecties met relatief korte strings (d.i. een relatief laag aantal deelstrings na splitsing) het aantal gefilterde koppels aanzienlijk is.

Uiteraard betekent dit voor bepaalde datacollecties ook dat een aantal coreferente koppels worden gefilterd en onterecht als niet-coreferent worden behan-

deld. Hoewel dit voor de meeste datacollecties niet zo is, merken we toch een maximum van 8.57% bij datacollectie 'univ'. Daarom verifiëren we de invloed van uitwisseling van kennis op de maximale f-waarde. Tabel 6.11 toont een vergelijking van maximale f-waarden in het geval van een parameterbepaling zonder trainingscollecties. Deze resultaten leren ons dat het niet gebruiken van kennisuitwisseling doorgaans leidt tot een lagere maximale f-waarde. De verklaring hiervoor is dat het niet gebruiken van kennisuitwisseling in verhouding meer valse positieven introduceert dan echte positieven. Dit doet ons besluiten dat het gebruik van kennisuitwisseling een significante verbetering biedt ten opzichte van de standaardaanpak.

naam	**met**	**zonder**
people	**0.9890**	0.9677
bird1	**0.9217**	**0.9217**
bird2	**0.8815**	**0.8815**
bird3	**0.9474**	**0.9474**
bird4	**0.9254**	0.9224
game	**0.7381**	0.6750
park	**0.9093**	**0.9093**
animal	**0.8787**	0.6628
census	**0.1789**	**0.1789**
univ	**0.8281**	**0.8281**
streets5	**0.9474**	**0.9474**
constraint	**0.9262**	**0.9262**
face	**0.9340**	**0.9340**
reasoning	**0.8811**	**0.8811**
reinforcement	**0.4615**	**0.4615**
restaurant	**0.8192**	**0.8192**

Tabel 6.11: Maximale f-waarden: met en zonder uitwisseling van kennis

De invloed van E_S Naast het al dan niet gebruiken van kennisuitwisseling willen we ook de invloed van de één-niveau methode onderzoeken. Zoals bij het begin van deze sectie is aangegeven, is bij alle experimenten verondersteld dat E_S het strenge toekenningsmodel gebruikt voor de generatie van possibiliteiten. Inspectie van de datacollecties leert dat een aantal datacollecties (in het bijzonder 'census') een ongewoon hoge foutenlast vertonen. Laat ons daarom bestuderen wat er gebeurt als we een soepeler toekenningsmodel vooropstellen, zoals het standaard toekenningsmodel (Tabel 6.4). Opnieuw gaan we bij de vergelijking uit van een parameterbepaling zonder trainingscollectie. De resultaten van deze vergelijking worden getoond in Tabel 6.12. De beste resultaten worden opnieuw in vet gezet. Een opvallend resultaat is de zeer sterke verbetering in maximale f-waarde voor datacollectie 'census'. Wanneer we dit vergelijken met de resultaten van eerdere experimenten stellen we vast dat enkel Levenshtein een betere maximale f-waarde haalt voor deze datacol-

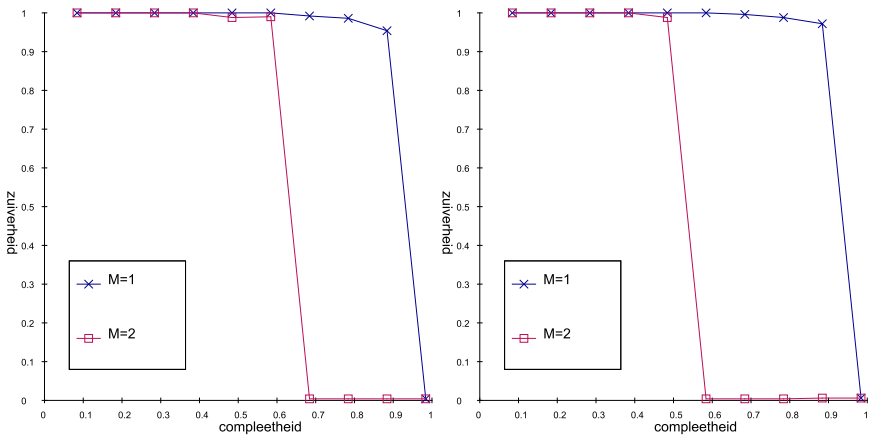
lectie. Onze twee-niveau aanpak doet hier echter beter dan SoftTFIDF. Voor de datacollecties met een relatief hoog aantal deelstrings na splitsing stellen we vast dat het minder strenge foutenmodel een negatieve invloed heeft op de maximale f-waarde, behalve voor 'reinforcement'. Dit is te wijten aan het feit dat de parameterbepaling voor deze datacollectie niet optimaal verloopt, d.i. α wordt veel te hoog geschat, waardoor de resulterende kwantorfunctie geen goede resultaten biedt. Een opmerkelijke vaststelling is echter dat het strenge foutenmodel voor vele datacollecties een goed resultaat vertoont, ondanks de strenge voorwaarden voor toekenning van possibiliteit voor coreferentie. Dit betekent dat het strenge foutenmodel in essentie volstaat om deelstrings te vergelijken, tenzij een zeer groot aantal fouten aanwezig is, zoals het geval bij 'census'. Dat dergelijke strenge toewijzing volstaat, motiveren we door te stellen dat enerzijds fouten meestal optreden in een minderheid van de deelstrings en anderzijds de kwantorfunctie een zekere marge biedt om deelstrings fout te beoordelen. Het kiezen voor het strenge toekenningsmodel heeft overigens ook een belangrijk voordeel met betrekking tot complexiteit. Immers, wanneer een deelverschil van het type prefix of het type fout wordt aangetroffen, kan de vergelijking van deelstrings vroegtijdig worden afgebroken. Anders gezegd, door het strenge toekenningsmodel te gebruiken hoeven de deelstrings niet volledig vergeleken te worden om tot een toekenning van possibiliteit te komen.

naam	**streng**	**standaard**
people	**0.9890**	**0.9890**
bird1	**0.9217**	**0.9217**
bird2	**0.8815**	**0.8815**
bird3	**0.9474**	**0.9474**
bird4	**0.9254**	**0.9254**
game	**0.7381**	**0.7381**
park	**0.9093**	0.9020
animal	**0.8787**	0.6875
census	0.1789	**0.7034**
univ	**0.8281**	**0.8281**
streets5	**0.9474**	**0.9474**
constraint	**0.9262**	0.9088
face	**0.9340**	0.9068
reasoning	**0.8811**	0.8632
reinforcement	0.4615	**0.5680**
restaurant	**0.8192**	**0.8192**

Tabel 6.12: Maximale f-waarden: E_S met streng en standaard toekenningsmodel

Meervoudige kwantificatie In Hoofdstuk 4 is het concept van meervoudige kwantificatie besproken. Hierbij wordt voor het zoeken naar coreferente collecties een kwantorfunctie bepaald na observatie van de collecties die worden vergeleken. Dit principe is toepasbaar in de context van twee-niveau verge-

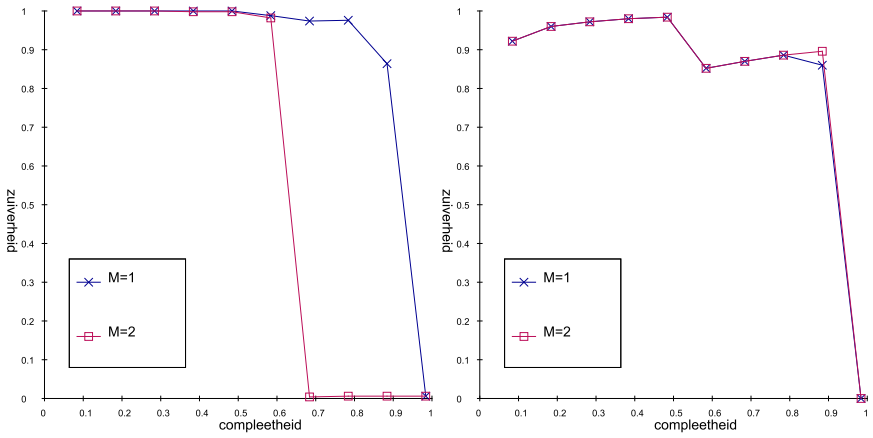
lijking van strings (Sectie 6.5). We voeren een experiment uit waarbij twee kwantorfuncties worden beschouwd. Dit doen we door eerst de parameters van één kwantorfunctie te bepalen zonder trainingsdata. Vervolgens kiezen we een getal b_1 zodat $[0, b_1[$ en $[b_1, 1]$ samen een partitie van het eenheidsinterval vormen. De berekende kwantorfunctie koppelen we aan het interval $[b_1, 1]$ en we koppelen een naïeve kwantorfunctie met $\alpha = 1$ aan het interval $[0, b_1[$. Een probleem dat zich dan stelt is de bepaling van b_1, gelet op het feit dat we niet beschikken over trainingsdata. Uit experimenten met trainingsdata is gebleken dat de waarde voor b_1 sterk invariabel is voor datacollecties met gelijkaardige karakteristieken, d.i. een gelijkaardig aantal deelstrings na splitsing. Uit deze experimenten hebben we afgeleid dat de waarde $b_1 = 0.9375$ een goede keuze is. Door aan het interval $[0, b_1[$ een vaste evaluator met $\alpha = 1$ te koppelen, wordt de toekenning van possibiliteiten verstrengd voor een deel van de koppels. Dit betekent dat we niet kunnen verwachten dat de invoering van meerdere kwantorfuncties zal leiden tot meer coreferente koppels die als dusdanig herkend worden. Het gebruik van een extra kwantorfunctie leidt met andere woorden a priori niet tot een daling van het aantal valse negatieven. Wel kunnen we verwachten dat het aantal valse positieven daalt. Gelet op het feit dat $|C| << |D|$ wil dit zeggen dat we zelden een positieve tendens in de maximale f-waarde zullen vaststellen. Om die reden wordt het effect van een bijkomende kwantorfunctie beter onderzocht aan de hand van het uitzetten van de *zuiverheid* ten opzichte van de *compleetheid*.



Figuur 6.9: Bijkomende kwantorfunctie: 'constraint' (links) en 'face' (rechts)

Voor de datacollecties waarbij een verschil wordt waargenomen in het verloop van *zuiverheid* ten opzichte van *compleetheid*, wordt dit verloop getoond in Figuren 6.9 tot 6.12. Hierbij duidt $m = 1$ op het gebruik van één klasse (d.i. enkelvoudige kwantificatie) en duidt $m = 2$ op het gebruik van twee klassen. Eerst en vooral valt op dat voor datacollecties 'constraint', 'face' en 'reasoning' de *zuiverheid* voor hogere waarden van *compleetheid* sterk daalt. Dit is niet

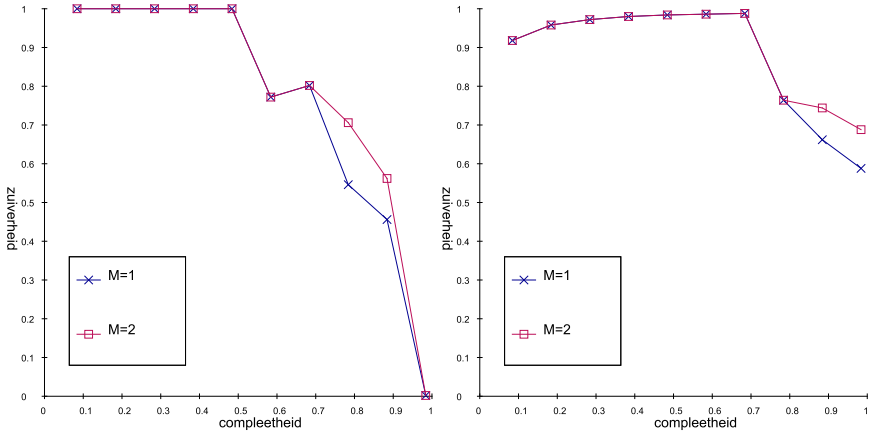
verwonderlijk gezien het feit dat voor deze datacollecties heel veel stringkoppels een kardinaliteitsverhouding hebben die in het interval $[0, b_1]$ valt. Deze strings worden behandeld met een kwantorfunctie waarvoor $\alpha = 1$ en het is reeds gebleken uit eerdere experimenten dat $\alpha = 1$ geen goede keuze is voor datacollecties waarbij een groot aantal deelstrings na splitsing wordt aangetroffen. Een mogelijke oplossing zou zijn om b_1 te verlagen voor datacollecties waarbij een groot aantal deelstrings overblijft na splitsing. We hebben echter vastgesteld dat een dergelijke ingreep leidt tot een neutralisatie van de invloed van de bijkomende kwantorfunctie. Een bijkomende kwantorfunctie is bijgevolg niet nuttig wanneer het aantal deelstrings na splitsing hoog is. Deze vaststelling wordt bevestigd door de redenering dat voor datacollecties met een groot aantal deelstrings na splitsing, het verschil tussen coreferente en niet-coreferente strings goed kan worden bepaald aan de hand van één enkele kwantorfunctie. Dit zal typisch minder makkelijk worden naarmate het aantal deelstrings daalt.



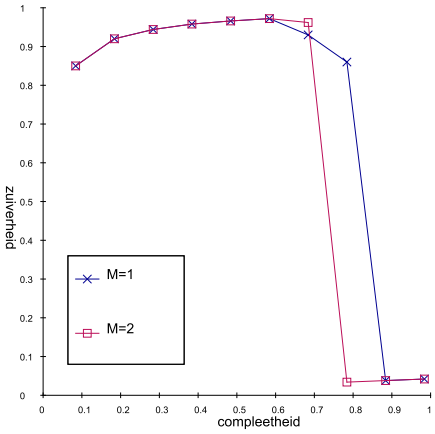
Figuur 6.10: Bijkomende kwantorfunctie: 'reasoning' (links) en 'animal' (rechts)

Voor de datacollecties met een kleiner aantal deelstrings stellen we vast dat het toevoegen van een extra kwantorfunctie inderdaad een percentage valse positieven weghaalt, wat tot gevolg heeft dat de staart van de *zuiverheid-compleetheidcurve* hoger komt te liggen, d.i. een verschuiving naar de rechterbovenhoek. Bij datacollectie 'univ' stellen we vast dat de toevoeging van een kwantorfunctie enkele valse negatieven introduceert, d.i. coreferente koppels die niet als dusdanig worden herkend en de possibilistische waarheidswaarde $(0, 1)$ toegewezen krijgen. Dit valt af te leiden uit het verschil in *zuiverheid* voor *compleetheid* 0.8 in Figuur 6.12.

Kardinaliteitsrestricties Tot slot van deze reeks experimenten willen we onderzoeken wat de invloed is van herstel van inconsistenties (Hoofdstuk 5). Hiervoor bestuderen we de invloed van het invoeren van kardinaliteitsrestric-



Figuur 6.11: Bijkomende kwantorfunctie: 'game' (links) en 'restaurant' (rechts)



Figuur 6.12: Bijkomende kwantorfunctie: 'univ'

ties. Voor 11 van de 16 datacollecties beschikken we namelijk over twee verschillende lijsten van objecten. Dit zijn alle datacollecties behalve 'univ', 'constraint', 'face', 'reasoning' en 'reinforcement'. Onder de veronderstelling dat elke lijst bestaat uit objecten die niet onderling coreferent zijn, is het mogelijk een restrictie te plaatsen op de partitieklassen die worden gevormd (Hoofdstuk 5). Deze restricties uiten zich door te vereisen dat elk object uit de eerste lijst met maximaal één object uit de tweede lijst coreferent kan zijn. In de praktijk wil dit zeggen dat we voor alle mogelijke koppels (s, t) de possibilistische waarheidswaarde $E^*_{\mathcal{S}}(s, t)$ berekenen. Daarna construeren we een één-op-één afbeelding tussen objecten uit de beide lijsten door gebruik te maken van Algoritme 4.1 (Hoofdstuk 4). We gaan opnieuw uit van een aanpak waarbij parameters bepaald worden zonder trainingsdata. Een vergelijking tussen de maximale

f-waarden wordt getoond in Tabel 6.13. De resultaten tonen een opmerkelijke verbetering in de maximale f-waarde voor heel wat datacollecties. Dit betekent dan ook dat het afdwingen van restricties op de kardinaliteit in deze situaties een sterke verbetering teweegbrengt in het zoeken naar coreferente objecten. De opmerkelijk lagere maximale f-waarde voor datacollectie 'animal' komt voort uit het aanwezig zijn van coreferente objecten binnen eenzelfde lijst. De één-op-één veronderstelling gaat voor deze datacollectie bijgevolg niet op, hetgeen wordt weerspiegeld in de resultaten.

naam	$E^*_{\mathcal{S}}$	$E^*_{\mathcal{S}}$ **met restricties**
people	0.9890	**1.0000**
bird1	0.9217	**0.9474**
bird2	0.8815	**0.9398**
bird3	**0.9474**	**0.9474**
bird4	0.9254	**0.9442**
game	0.7381	**0.7500**
park	0.9093	**0.9395**
animal	**0.8787**	0.7775
census	**0.1789**	**0.1793**
streets5	**0.9474**	**0.9474**
restaurant	0.8192	**0.8774**

Tabel 6.13: Maximale f-waarden: invloed van kardinaliteitsrestricties

6.9 Conclusie

In dit hoofdstuk is het coreferentieprobleem bestudeerd in het geval van karakterstrings (ook strings genoemd). Onze aanpak vertrekt van een formalisatie van het concept van een string, waarbij een aantal operatoren worden ingevoerd. Vervolgens introduceren we een nieuwe aanpak voor het vergelijken van strings die vertrekt van een één-niveau evaluator. Een dergelijke evaluator behandelt een string als een sequentie van karakters en evalueert enkel en alleen op basis van deze structuur. We bespreken hoe een evaluator op basis van deelverschillen tussen strings komt tot de toekenning van possibiliteit. Een voordeel van onze één-niveau evaluator is zijn lage complexiteit in vergelijking met bestaande karaktergebaseerde methoden. Vervolgens beschouwen we een twee-niveau evaluator door strings op te splitsen tot een multiverzameling van deelstrings. Een dergelijke collectie van deelstrings kan worden vergeleken door gebruik te maken van de resultaten uit Hoofdstuk 4. In het possibilistisch twee-niveau raamwerk worden enkele nieuwe en interessante ideeën onderzocht. Een eerste idee is het uitwisselen van kennis tussen niveaus, wat betekent dat onzekerheid over coreferentie iteratief wordt berekend zodat in opeenvolgende iteraties de combinatiefunctie en de één-niveau evaluator variëren. We geven een voorbeeld van dergelijke iteratieve berekening en tonen aan dit leidt

tot het filteren van strings. Daarnaast worden ook het gebruik van meerdere kwantorfuncties, het gebruik van een frequentiefilter en het belang van de splitsingsfunctie onderzocht. Er worden methoden gegeven voor het bepalen van de parameters van een kwantorfunctie, enerzijds op basis van trainingsdata en anderzijds zonder trainingsdata. Ten slotte worden al deze technieken vergeleken met geavanceerde methoden uit de literatuur. Uit deze experimenten blijkt dat de één-niveau evaluator aanzienlijk sneller werkt dan bestaande karaktergebaseerde methoden, maar dat hij als alleenstaande evaluator slechts goed presteert wanneer de verschillen tussen coreferente strings beperkt zijn. De twee-niveau evaluator toont op vlak van accuraatheid (gemeten aan de hand van de maximale f-waarde) heel wat betere resultaten, voornamelijk wanneer de kwantorfunctie wordt aangepast aan de gegeven datacollectie. Onze experimenten tonen dat dit het geval is zowel met als zonder het gebruik van trainingsdata. Het is gebleken dat het uitwisselen van kennis tussen evaluatoren een positieve invloed heeft op de complexiteit zonder de accuraatheid aan te tasten. De invloed van meervoudige kwantificatie is voornamelijk te merken bij datacollecties met relatief korte strings. Hierbij wordt vastgesteld dat meervoudige kwantificatie in staat is de *zuiverheid* voor de hoogste waarden van *compleetheid* te verbeteren voor bepaalde datacollecties. Voor datacollecties met relatief lange strings heeft meervoudige kwantificatie een negatieve invloed. Tot slot is vastgesteld dat het afdwingen van kardinaliteitsrestricties in veel gevallen leidt tot hogere maximale f-waarden.

Hoofdstuk 7

Evaluatoren voor complexe objecten

7.1 Inleiding

In voorgaande hoofdstukken zijn evaluatoren gedefinieerd en is onderzocht wat hun eigenschappen zijn. Er is een onderscheid gemaakt tussen semantische en syntactische evaluatoren en beide klassen van evaluatoren zijn onderzocht. In het algemene geval is het probleem van moeilijk-meetbaarheid onderzocht. Bij syntactische evaluatoren is onderzocht hoe collecties vergeleken kunnen worden en hoe deze resultaten gebruikt kunnen worden voor de bepaling van onzekerheid over coreferentie in het geval van strings. Er is aangetoond dat wanneer verschillende actoren onzekerheid uitdrukken over de waarheidswaarde van één propositie p, een raamwerk van conditionele necessiteit de mogelijkheid biedt om deze kennis te combineren. Daarnaast is ook bestudeerd hoe een evaluator kan worden gebruikt om een relatie te construeren die voldoet aan de vereisten van de coreferentierelatie $\leftrightarrow$. Een dergelijke relatie kan worden gebruikt als benadering voor $\leftrightarrow$ en biedt als dusdanig een oplossing voor het coreferentieprobleem. Echter, tot hier toe is de constructie van evaluatoren voornamelijk gericht op het geval van atomaire objecten (collecties[1], strings ...). In dit hoofdstuk willen we al deze resultaten samenbrengen in een raamwerk voor de bepaling van coreferentie van complexe objecten. Eerst zal in Sectie 7.2 een overzicht worden gegeven van bestaande methoden. Daarna wordt in Sectie 7.3 het possibilistische raamwerk voor coreferentiebepaling van complexe objecten voorgesteld. Hierbij worden drie stappen besproken. Eerst wordt het belang van de keuze voor evaluatoren over deeluniversa toegelicht. Daarna wordt besproken hoe parameters van deze evaluatoren bepaald kunnen worden. Ten slotte wordt onderzocht hoe conditionele necessiteit bepaald kan worden. In

[1] Hoewel een collectie volgens de ODMG standaard niet atomair is, is het universum van dergelijke collecties niet te schrijven als een combinatie van andere universa. Bijgevolg vallen collecties in deze thesis onder de noemer 'atomaire' objecten.

Sectie 7.4 worden enkele experimentele resultaten gerapporteerd die aantonen hoe onze aanpak zich verhoudt tot bestaande methoden. Sectie 7.5 biedt een overzicht van de belangrijkste resultaten en bevindingen uit dit hoofdstuk.

7.2 Overzicht van de literatuur

Het zoeken naar koppels van coreferente complexe objecten wordt in de literatuur vaak het *record linkage* probleem genoemd. Deze benaming wordt toegeschreven aan Dunn [103], die voor het ontstaan van moderne opslagsystemen voor data in 1946 reeds wijst op het belang van wat hij noemt het *boek van het leven*. Daarmee duidt hij de volledige verzameling van medische gegevens over één persoon aan en schetst hij de noodzaak aan een systeem waarmee deze gegevens eenvoudig aan elkaar te koppelen zijn. De eerste pogingen om een theoretische onderbouw aan dit probleem te geven, worden ondernomen door Newcombe [93, 104] die gebruik maakt van de Soundex-code voor het vergelijken van velden[2] en die duidt op het verschil in discriminerende kracht tussen velden. Fellegi en Sunter bouwen verder op de aanpak van Newcombe en geven in [42] met het FS-model voor het eerst een volwaardige, probabilistische oplossing voor het coreferentieprobleem. Het FS-model vertrekt van twee collecties van records (d.z. objecten) A en B die vergeleken moeten worden. Elk koppel $(a, b) \in A \times B$ moet daarom worden geclassificeerd als een *match* (M) of een *non-match* (U). Fellegi en Sunter veronderstellen dat elk record bestaat uit n velden[2]. Omdat deze velden dezelfde zijn voor alle records onder vergelijking, laten Fellegi en Sunter elk koppel (a, b) corresponderen met een *overeenkomstpatroon* $\mathbf{x} \in [0, 1]^n$ waarbij $\mathbf{x}_i$ de overeenkomst voor het i^{de} veld voorstelt. Fellegi en Sunter beschrijven een methode om een probabilistische beslissingsregel op te stellen die *overeenkomstpatronen* afbeeldt op de classificatieruimte $\{M, U\}$. Hun methode steunt op de veronderstelling dat klassen M en U aanleiding geven tot patronen met een verschillende verdeling. De beslissingsregel van Fellegi en Sunter kan worden geschreven als:

$$(a, b) \in \begin{cases} M & \textbf{als} \quad \frac{\Pr(\mathbf{x}|M)}{\Pr(\mathbf{x}|U)} \geq \frac{\Pr(U)}{\Pr(M)} \\ U & \textbf{anders}. \end{cases} \tag{7.1}$$

Deze beslissingsregel kan worden gezien als het zoeken naar een optimale partitie van een vaste vergelijkingsruimte $X = [0, 1]^n$ in twee partitieklassen. Met vast wordt hier bedoeld dat de generatie van patronen $\mathbf{x}$ een vast proces is, eerder dan een aanpasbare parameter van het vergelijkingssysteem. De nadruk hierop is niet toevallig, aangezien onze possibilistische aanpak op dit vlak zal verschillen van de probabilistische aanpak van Fellegi en Sunter. Enerzijds zullen we aandacht besteden aan de keuze van evaluatoren voor de deeluniversa van een complex universum. Anderzijds is het aanpassen van parameters van deelevaluatoren aan een gegeven datacollectie een belangrijk onderdeel van onze possibilistische aanpak.

[2] De term 'veld' slaat hier op een deelobject die een eigenschap van een entiteit beschrijft.

Wanneer een verzameling van patronen $\mathbf{x}$ met bijhorende klasse beschikbaar is, kunnen de conditionele waarschijnlijkheden ($\Pr(\mathbf{x}|M)$ en $\Pr(\mathbf{x}|U)$) en de *priors* ($\Pr(U)$ en $\Pr(M)$) worden aangeleerd. Het is echter geweten dat het opbouwen van een trainingscollectie een dure operatie is. Dit is te wijten aan het typisch zeer kleine percentage van coreferente koppels, waardoor het voorzien van koppels met klasselabel M veel inspanning vergt. Om die reden wordt dan ook veel aandacht besteed aan technieken waarbij een trainingscollectie niet nodig is. Verschillende methoden voor schatting zonder training zijn voorgesteld in de literatuur. Jaro veronderstelt in [105] conditionele onafhankelijkheid tussen de verschillende velden van records. Dit betekent dat er geldt:

$$\Pr(\mathbf{x}|M) = \prod_{i=1}^{n} \Pr(\mathbf{x}_i|M) \tag{7.2}$$

$$\Pr(\mathbf{x}|U) = \prod_{i=1}^{n} \Pr(\mathbf{x}_i|U)\,. \tag{7.3}$$

Verder gebruikt Jaro binaire *overeenkomstpatronen* (d.i. $\mathbf{x}_i \in \{0,1\}$) en een blokmethode om een steekproef uit $A \times B$ te trekken. Een dergelijke blokmethode sorteert de koppels in $A \times B$ en selecteert koppels door een vast venster over de gesorteerde koppels te beschouwen. Jaro kiest zijn blokmethode zo dat koppels in de steekproef met grote waarschijnlijkheid tot de klasse M behoren. De verkregen steekproef kan worden gebruikt om de onbekende waarschijnlijkheden te schatten met behulp van het EM-algoritme (Expectation Maximization) [106]. In de meer algemene aanpak van Winkler wordt de veronderstelling van conditionele onafhankelijkheid weggelaten. In de plaats stelt Winkler een reeks voorwaarden voor, die samen minder streng zijn dan conditionele onafhankelijkheid [107]. Winkler past het EM-algoritme aan zodat het met de minder strenge voorwaarden rekening houdt. Winkler veronderstelt een oneindige, niet-aftelbare vergelijkingsruimte X, maar gebruikt een handmatige discretisatie van de *overeenkomstpatronen* om het gebruik van zijn versie van het EM-algoritme mogelijk te maken [108]. Hij beschrijft echter nergens wat de impact is van zijn handmatig gekozen grenzen. Een dergelijke discretisatie willen we in onze aanpak vermijden. Een belangrijk verschil tussen onze aanpak en de probabilistische modellen zal bijgevolg liggen in de aanpak van vergelijking van deelobjecten.

Hoewel het model van Fellegi en Sunter terecht als een historische mijlpaal kan worden gezien, zijn een aantal alternatieve methoden het vermelden waard. In [109] wordt door Du Bois een model geformuleerd waarin expliciet aandacht wordt besteed aan ontbrekende data. Du Bois breidt hiervoor het probabilistische model van Fellegi en Sunter verder uit. Er zijn ook alternatieve methoden te vinden die niet vertrekken van het model van Fellegi en Sunter. Zo zijn er bijvoorbeeld regelgebaseerde systemen voorgesteld die gebruik maken van beslissingsbomen [110, 111, 102, 112]. In het kader van enerzijds het objectgeoriënteerd paradigma voor moderne programmeertalen en anderzijds XML voor

de voorstelling van data op het WWW, is heel wat recent onderzoek gericht op meer geavanceerde datastructuren om complexe objecten voor te stellen [113, 71, 114]. De noodzaak aan het omgaan met dergelijke geavanceerde datastructuren wordt in onze aanpak weerspiegeld door de algemene definitie van objecten (Hoofdstuk 2).

7.3 Een possibilistisch raamwerk

Rekening houdend met de vermelde methoden willen we hier een possibilistisch raamwerk voor coreferentie van complexe objecten introduceren. Daarbij willen we de volgende voordelen nastreven. Ten eerste, willen we in onze aanpak het belang van vergelijking op het niveau van eigenschappen van objecten niet onderschatten. In alle bestaande methoden wordt meestal een vaste functie verondersteld voor vergelijking op dit niveau. In onze aanpak willen we aan de keuze van deze functies voldoende aandacht schenken. Ten tweede willen we vermijden dat voor het bepalen van conditionele necessiteit een discretisatiestap nodig is, net zoals dit nodig is bij de schatting van conditionele waarschijnlijkheid in de aanpak van Winkler [107, 108]. We zullen dit doen door te steunen op enerzijds de eigenschappen van conditionele necessiteit (Hoofdstuk 3) en anderzijds de eigenschappen van beslissingsmodellen (Hoofdstuk 5). Ten derde willen we een aanpak vooropstellen die weinig veronderstellingen maakt over de te vergelijken objecten en bijgevolg algemeen toepasbaar is op problemen van een uiteenlopende aard. Als laatste bemerken we dat het gebruik van possibiliteiten een elegant en natuurlijk model biedt voor het omgaan met niet-meetbare en moeilijk-meetbare eigenschappen (Hoofdstuk 2).

7.3.1 Definities

Beschouw een universum $O = U_1 \times ... \times U_n$ van complexe objecten. Voor elke twee objecten o_1 en o_2 uit dit universum is de coreferentiële propositie gegeven (Definitie 2.11) als:

$$p_{(o_1,o_2)} = \text{“}o_1 \text{ en } o_2 \text{ zijn coreferent”}.$$

We weten uit Hoofdstuk 2 dat elk complex object o bestaat uit n deelobjecten die elk een eigenschap van een entiteit beschrijven. Coreferentie van deelobjecten wordt uitgedrukt door proposities van de vorm:

$$p_{(o_1,o_2)_i} = \text{“proj}_i(o_1) \text{ en proj}_i(o_2) \text{ zijn coreferent”}.$$

Laat ons nu veronderstellen dat een actor $\mathcal{A}_i$ gebruik maakt van het volgende inferentiemechanisme:

$$\left(\mathcal{A}_i \rightsquigarrow \left(p_{(o_1,o_2)_i} = T\right)\right) \vdash \left(\mathcal{A}_i \rightsquigarrow \left(p_{(o_1,o_2)} = T\right)\right) \tag{7.4}$$

$$\left(\mathcal{A}_i \rightsquigarrow \left(p_{(o_1,o_2)_i} = F\right)\right) \vdash \left(\mathcal{A}_i \rightsquigarrow \left(p_{(o_1,o_2)} = F\right)\right). \tag{7.5}$$

Een actor $\mathcal{A}_i$ baseert zijn postulaten over de waarheidswaarde van $p_{(o_1,o_2)}$ op de kennis die hij heeft over de waarheidswaarde van $p_{(o_1,o_2)_i}$. Elke eigenschap

wordt bijgevolg gezien als een stukje evidentie voor het (niet) coreferent zijn van complexe objecten. Anders gezegd, door het afzonderlijk beschouwen van de eigenschappen verkrijgen we n postulaten over de waarheidswaarde van $p_{(o_1,o_2)}$. Wanneer de meetprocessen voor de eigenschappen perfect zijn zouden deze postulaten allemaal dezelfde moeten zijn, d.i. allemaal $p_{(o_1,o_2)} = T$ of allemaal $p_{(o_1,o_2)} = F$. Bij imperfecte meetprocessen kunnen er conflicten bestaan tussen de postulaten over de waarheidswaarde van p.

Als elk deeluniversum U_i gelijk is aan $\mathcal{S}$, bestaat een naïeve aanpak erin om voor elk object o de string s_o te beschouwen zodat:

$$s_o = \bigoplus_{i=1}^{n} \text{proj}_i(o). \tag{7.6}$$

Coreferentie van complexe objecten wordt in dit geval herleid tot coreferentie van strings. Dat deze aanpak niet noodzakelijk leidt tot slechte resultaten, wordt aangetoond in Sectie 6.8. Verschillende datacollecties gebruikt voor de experimenten in Hoofdstuk 6 bevatten namelijk strings die complexe objecten voorstellen. Een dergelijke aanpak kan nuttig zijn als complexe objecten enkel beschikbaar zijn als een string. Er bestaan technieken die strings omzetten naar complexe objecten [115], maar dergelijke technieken zijn niet foutloos, zodat de aanpak op basis van stringvergelijking nuttig kan zijn. Een nadeel van het werken met strings is dat er geen rekening wordt gehouden met het mogelijke verschil in discriminerende kracht tussen de eigenschappen van een entiteit [104]. Wanneer de structuur van een complex object gekend is, bestaat er een betere oplossing die rekening houdt met dergelijke verschillen. Als de waarheidswaarden van één of meerdere actoren postuleren dat $p_{(o_1,o_2)}$ waar is, zal onze possibilistische aanpak de verschillen in discriminerende kracht in de zin van Newcombe [104] uitdrukken als conditionele necessiteit voor het waar zijn van $p_{(o_1,o_2)}$. Dit leidt ons naar de volgende definitie.

Definitie 7.1 (Evaluator voor complexe objecten)
Gegeven een universum van complexe objecten $O = U_1 \times ... \times U_n$. Een evaluator voor complexe objecten is gedefinieerd als:

$$\begin{array}{rcl} E_O : O^2 & \rightarrow & \mathcal{F}(\mathbb{B}) \\ (o_1, o_2) & \mapsto & S_{\gamma^{T,F}}\left(\left\{E_{U_i}\left(\text{proj}_i(o_1), \text{proj}_i(o_2)\right) | i \in \{1, ..., n\}\right\}\right). \end{array} \tag{7.7}$$

Uit Definitie 7.1 kan worden afgeleid dat een evaluator E_O bestaat uit deelevaluatoren E_{U_i} en een combinatiefunctie $S_{\gamma^{T,F}}$. We merken op dat een evaluator E_O steeds reflexief en symmetrisch is. Een evaluator E_O is niet noodzakelijk sterk reflexief en is niet noodzakelijk transitief. Met betrekking tot sterke reflexiviteit geldt de volgende stelling.

Stelling 7.1
Een evaluator voor complexe objecten E_O is sterk reflexief als er minstens één deelevaluator E_{U_i} sterk reflexief is en als er geldt dat $S_{\gamma^{T,F}} = \widetilde{\wedge}$.

Bewijs. Voor twee verschillende objecten o_1 en o_2 geldt er dat:

$$\exists i \in \{1, ..., n\} : E_{U_i}\big(\text{proj}_i(o_1), \text{proj}_i(o_2)\big) \neq (1, 0). \tag{7.8}$$

Gelet op $S_{\gamma^{T,F}} = \widetilde{\wedge}$ geldt er dan dat:

$$E_O(o_1, o_2) \neq (1, 0). \tag{7.9}$$

□

Met betrekking tot transitiviteit geldt de volgende stelling.

Stelling 7.2
Een evaluator voor complexe objecten E_O is transitief als alle deelevaluatoren E_{U_i} transitief zijn en als $S_{\gamma^{T,F}} = \widetilde{\wedge}$.

Bewijs. Beschouw drie willekeurige objecten o_1, o_2 en o_3. We onderscheiden drie gevallen.

(1) In het eerste geval veronderstellen we dat er geldt:

$$E_O(o_1, o_2) = (1, a) \wedge E_O(o_2, o_3) = (1, b) \tag{7.10}$$

en bovendien veronderstellen we dat $a < b$. Dit betekent dat:

$$(1, a) > (1, b). \tag{7.11}$$

Gelet op het feit dat $S_{\gamma^{T,F}} = \widetilde{\wedge}$, weten we dat er een $k \in \{1, ..., n\}$ bestaat waarvoor er geldt dat:

$$E_{U_k}\Big(\text{proj}_k(o_2), \text{proj}_k(o_3)\Big) = (1, b) \tag{7.12}$$

en:

$$E_{U_k}\Big(\text{proj}_k(o_1), \text{proj}_k(o_2)\Big) > (1, b). \tag{7.13}$$

Door de transitiviteit van E_{U_k} geldt er dan dat:

$$E_{U_k}\Big(\text{proj}_k(o_1), \text{proj}_k(o_3)\Big) = (1, b). \tag{7.14}$$

Bovendien weten we dat:

$$\forall i \in \{1, ..., n\} : E_{U_i}\Big(\text{proj}_i(o_1), \text{proj}_i(o_2)\Big) \geq (1, b) \tag{7.15}$$

en

$$\forall i \in \{1, ..., n\} : E_{U_i}\Big(\text{proj}_i(o_2), \text{proj}_i(o_3)\Big) \geq (1, b) \tag{7.16}$$

zodat er omwille van de transitiviteit van alle deelevaluatoren geldt dat:

$$\forall i \in \{1, ..., n\} : E_{U_i}\Big(\text{proj}_i(o_1), \text{proj}_i(o_3)\Big) \geq (1, b). \tag{7.17}$$

Gelet op $S_{\gamma^{T,F}} = \widetilde{\wedge}$ weten we bijgevolg dat er geldt:

$$E_O(o_1, o_3) = (1, b). \tag{7.18}$$

(2) In het tweede geval veronderstellen we dat er geldt:

$$E_O(o_1, o_2) = (1, a) \wedge E_O(o_2, o_3) = (b, 1) \tag{7.19}$$

en bovendien veronderstellen we dat $a < b$. Indien er voor een $i \in \{1, ..., n\}$ geldt dat:

$$E_{U_i}\Big(\text{proj}_i(o_2), \text{proj}_i(o_3)\Big) = (1, c) \tag{7.20}$$

dan weten we omwille van de transitiviteit van E_{U_i} dat er een $d \in [0, 1]$ bestaat zodat:

$$E_{U_i}\Big(\text{proj}_i(o_1), \text{proj}_i(o_3)\Big) = (1, d). \tag{7.21}$$

Indien er voor een $i \in \{1, ..., n\}$ geldt dat:

$$E_{U_i}\Big(\text{proj}_i(o_2), \text{proj}_i(o_3)\Big) = (c, 1) \tag{7.22}$$

dan moet er gelden dat:

$$c \geq b > a. \tag{7.23}$$

Omwille van de transitiviteit van E_{U_i} wil dit zeggen dat:

$$E_{U_i}\Big(\text{proj}_i(o_1), \text{proj}_i(o_3)\Big) = (c, 1) \tag{7.24}$$

Gelet op $S_{\gamma^{T,F}}$ moet er dan gelden dat:

$$E_O(o_1, o_3) = (b, 1). \tag{7.25}$$

(3) In het derde geval veronderstellen we dat er geldt:

$$E_O(o_1, o_2) = (1, a) \wedge E_O(o_2, o_3) = (b, 1) \tag{7.26}$$

en bovendien veronderstellen we dat $a \geq b$. Er moet dan gelden dat:

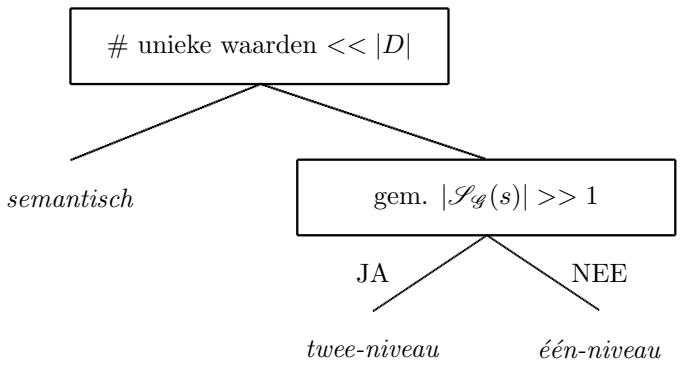
$$E_O(o_1, o_3) = (c, 1) \tag{7.27}$$

met $a \geq c$. Zoniet, is er een tegenspraak met het gegeven dat $E_O(o_2, o_3) = (b, 1)$. □

In wat volgt onderzoeken we een aanpak om een evaluator E_O voor een gegeven datacollectie D te construeren, zonder het gebruik van een trainingscollectie. Daarbij worden drie stappen beschouwd: selectie van deelevaluatoren, parameterbepaling van deelevaluatoren en bepaling van conditionele necessiteit. Deze stappen worden één voor één beschreven in de nu volgende deelsecties. Daarna lichten we het gebruik van beslissingsmodellen toe.

7.3.2 Selectie van deelevaluatoren

De constructie van een evaluator E_O vereist dat voor elk deeluniversum U_i een evaluator wordt voorzien. Deze stap lijkt bijzonder eenvoudig, maar het is onze vaststelling dat deze stap in de literatuur weinig aandacht krijgt. Nochtans heeft onze bespreking van stringvergelijking geleerd dat voor $\mathcal{S}$ heel wat uiteenlopende methoden bestaan, elk met specifieke veronderstellingen over het foutenmodel dat van toepassing is (Hoofdstuk 6). In onze possibilistische aanpak van het coreferentieprobleem hebben we in Hoofdstuk 2 een onderscheid gemaakt tussen verschillende oorzaken van onzekerheid. Hierbij aansluitend argumenteren we dan ook dat een deelevaluator het resultaat moet zijn van een weloverwogen keuze. De problematiek van niet-meetbaarheid en moeilijkmeetbaarheid van eigenschappen en daarbij aansluitend het geval van possibilistische databanken is reeds besproken in Hoofdstuk 2. In datzelfde hoofdstuk is ook het niet-éénduidig bepaald zijn van het resultaat van een meting besproken. Semantische evaluatie is daarbij vooropgesteld als mogelijke oplossing. De aanwezigheid van ruis kan worden aangepakt door een foutenmodel op te stellen dat de impact van ruis op objecten in kaart brengt. Een dergelijk foutenmodel kan dan door een syntactische evaluator worden gebruikt. Dit hebben we onderzocht in het geval van strings (Hoofdstuk 6) en dit heeft geleid tot de definitie van één-niveau en twee-niveau evaluatoren voor strings. De variabiliteit van entiteitsbeschrijvingen over de tijd kunnen we niet oplossen aan de hand van een evaluator. Tijdsvariabiliteit wordt in onze aanpak gezien als een fout (deel)object en kan worden aangepakt bij de bepaling van conditionele necessiteit. Door uit te drukken dat het coreferent zijn van een deel van de deelobjecten necessiteit geeft voor het coreferent zijn van de objecten, kunnen we rekening houden met deelobjecten van coreferente objecten die niet als coreferent bestempeld worden omwille van een fout of tijdsvariabiliteit.



Figuur 7.1: Keuzecriterium: deelevaluatoren voor $\mathcal{S}$

Voor elke oorzaak van onzekerheid die beschouwd is, bestaat er een oplossing in het possibilistische raamwerk. De vraag blijft echter hoe we automatisch

een bepaalde oorzaak kunnen detecteren. In een context zonder supervisie is dit bijzonder moeilijk en kan dit enkel op basis van een studie van de karakteristieken van objecten. In het geval waarbij $U_i = \mathcal{S}$ is het mogelijk enkele heuristische regels te formuleren, die zich laten samenvatten als een beslissingsboom (Figuur 7.1). Deze heuristiek steunt op het feit dat semantische evaluatoren nuttig zijn als het aantal verschillende deelobjecten voor één eigenschap binnen een datacollectie zeer beperkt is. Bij een syntactische evaluator wordt gekozen voor een één-niveau evaluator voor strings als het gemiddeld aantal deelstrings na splitsing zeer laag is. Zoniet wordt gekozen voor een twee-niveau evaluator. Uit experimenten blijkt dat een dergelijke heuristiek goede resultaten oplevert. In deze thesis leggen we ons voornamelijk toe op het geval van strings, maar de gevolgde redenering kan eveneens worden gevolgd voor andere atomaire universa.

7.3.3 Bepaling van parameters voor deelevaluatoren

Na het toekennen van een evaluator aan elk deeluniversum U_i moeten eventuele parameters van deze evaluatoren worden bepaald. Voor de bepaling van zijn parameters kan elke deelevaluator E_{U_i} gebruik maken van de i^{de} projectie van de datacollectie $D \subset O^2$:

$$\text{proj}_i(D) = \{(u_1, u_2) | \exists (o_1, o_2) \in D : \text{proj}_i(o_1) = u_1 \wedge \text{proj}_i(o_2) = u_2\}\,. \quad (7.28)$$

Door het gebruik van deze projecties kunnen de parameters van een tweeniveau evaluator voor strings worden bepaald, zoals beschreven in Hoofdstuk 6. Deze parameterbepaling steunt op de vaststelling dat de daling in *zuiverheid* voor variërende parameters evenredig is met de stijging in het aantal koppels waarvoor necessiteit voor coreferentie wordt gevonden. Hoewel in deze thesis de nadruk ligt op strings, wordt in lopend onderzoek bestudeerd of dit principe ook kan worden toegepast voor numerieke objecten. Voor sommige evaluatoren is het expliciet bepalen van de parameters niet nodig, gezien de algemeenheid van het foutenmodel. Zo zullen we voor een één-niveau evaluator steeds gebruik maken van het standaardtoekenningsmodel (Tabel 6.4), dat aanleiding geeft tot een standaardevaluator. Voor semantische evaluatoren zullen we de relatie R construeren als volgt. Beschouw voor een universum $O = U_1 \times ... \times U_n$ een datacollectie $D \subset O^2$. Als voor een welbepaalde U_i een evaluator E_{U_i} gegeven is, dan bepalen we de kandidaatverzameling $\widehat{C}_i$ als volgt:

$$\widehat{C}_i = \{(o_1, o_2) | (o_1, o_2) \in D \wedge \mathscr{B}\left(E_{U_i}(\text{proj}_i(o_1), \text{proj}_i(o_2))\right) = T\} \quad (7.29)$$

met $\mathscr{B}$ een maximaal tweewaardig beslissingsmodel. Anders gezegd, de verzameling $\widehat{C}_i$ bevat koppels (o_1, o_2) van complexe objecten waarvoor een necessiteit groter dan 0 bestaat dat de i^{de} deelobjecten coreferent zijn. Voor het complex universum $O = U_1 \times ... \times U_n$ beschouwen we de indexverzameling $I_O = \{1, ..., n\}$. Beschouw nu een verzameling $J \subset I_O$ zodat voor elke $j \in J$ geldt dat het universum U_j voorzien is van een syntactische evaluator, waarvoor

eventuele parameters zijn bepaald. Beschouw dan voor elke $i \notin J$ de binaire multirelatie $R_i \subset \mathcal{M}(U_i \times U_i)$, zodat er voor elk koppel $(u_1, u_2) \in U_i^2$ geldt:

$$\omega_{R_i}(u_1, u_2) = \sum_{j \in J} \left| \left\{ (o_1, o_2) | (o_1, o_2) \in \widehat{C}_j \wedge \text{proj}_i(o_1) = u_1 \wedge \text{proj}_i(o_2) = u_2 \right\} \right|. \tag{7.30}$$

Een semantische evaluator voor het deeluniversum U_i is dan gegeven als:

$$E_{(R_i)_k, U_i} \tag{7.31}$$

waarbij $(R_i)_k$ een k-snede is van de multirelatie R_i. Dit principe is ook aangehaald in Hoofdstuk 2. Een semantische evaluator wordt hierdoor dynamisch geconstrueerd op basis van de evidentie die wordt gegeven door syntactische evaluatoren.

7.3.4 Bepaling van conditionele necessiteit

De derde stap in de constructie van E_O is de bepaling van conditionele necessiteit. Newcombe haalt in [104] aan dat deelobjecten een verschillende discriminerende kracht kunnen hebben. In onze aanpak stellen we dat er aan elke mogelijke deelverzameling van actoren een bepaalde conditionele necessiteit wordt toegekend die het vertrouwen in die actoren voorstelt. Noteren we de verzameling van actoren als $A = \{\mathcal{A}_1, ..., \mathcal{A}_n\}$, dan betekent de bepaling van conditionele necessiteit dat we voor elke $Q \subseteq A$ moeten afleiden wat de necessiteit is dat $p_{(o_1,o_2)}$ waar (resp. vals) is als de actoren in Q postuleren dat $p_{(o_1,o_2)} = T$ (resp. $p_{(o_1,o_2)} = F$).

Bij het bepalen van conditionele necessiteit spelen twee aspecten een rol. Enerzijds is het zo dat actoren overeenkomend met verschillende eigenschappen een verschillende conditionele necessiteit impliceren. Dit strookt met de vroege intuïtie van Newcombe, die stelt dat het discriminerend vermogen van eigenschappen sterk kan uiteenlopen. Zo heeft bijvoorbeeld de voornaam van een persoon een sterker discriminerend vermogen dan zijn of haar geslacht. Anderzijds speelt de combinatie van eigenschappen een belangrijke rol. Een typisch verschijnsel bij het coreferentieprobleem is dat twee eigenschappen die elk afzonderlijk weinig conditionele necessiteit bieden, samen een hoge conditionele necessiteit kunnen bieden. In [116] hebben we aangetoond dat combinaties van eigenschappen met een hoge necessiteit typisch gevonden worden door gebruik te maken van de groeperingsfunctie λ (Hoofdstuk 2). Dit houdt in dat voor een groep van k actoren:

$$Q = \{\mathcal{A}_1, ..., \mathcal{A}_k\} \tag{7.32}$$

de conditionele necessiteit hoger is wanneer:

$$\lambda\{\text{lab}(U_1), ..., \text{lab}(U_k)\} = 1 \tag{7.33}$$

dan wanneer dit niet het geval is. Hierbij is lab een functie die aan elk universum een unieke naam koppelt. Een groeperingsfunctie kan bijgevolg bijzonder

nuttig zijn bij het construeren van conditionele necessiteit. Zelfs wanneer deze functie niet expliciet aanwezig is, bestaat ze meestal impliciet en kan ze handmatig worden opgesteld. We zullen in wat volgt een meer algemene aanpak beschrijven voor de bepaling van conditionele necessiteit. Deze aanpak kan echter steeds worden verfijnd door gebruik te maken van een groeperingsfunctie.

We zullen eerst enkele theoretische overwegingen maken die de algemeenheid van ons model benadrukken. Daarna zullen we een praktische aanpak voorstellen om conditionele necessiteit te bepalen. In Hoofdstuk 3 is aangehaald dat een bijzonder voordeel van de Sugeno-gebaseerde aanpak het bipolaire aspect van de combinatiefunctie is. In het bijzonder kunnen er verzamelingen van actoren $Q \subseteq A$ bestaan zodat:

$$\gamma^T(Q) = \gamma^F(Q) = 0. \tag{7.34}$$

Dit betekent dat de kennis die aangeboden wordt door actoren in Q geen bijdrage levert tot de kennis over de waarheidswaarde van p. In termen van coreferentie wil dit zeggen dat bepaalde (combinaties van) eigenschappen geen uitsluitsel kunnen bieden over het al dan niet coreferent zijn. Een dergelijke strategie kan voordelig zijn in situaties waar het vermijden van fouten bijzonder kritisch is: eerder dan het nemen van een beslissing over de waarheidswaarde, wordt expliciet aangeduid dat de waarheidswaarde niet kan worden bepaald. Dit gegeven sluit naadloos aan bij het principe van driewaardige beslissingsmodellen uit Hoofdstuk 5. In een driewaardig beslissingsmodel zal de possibilistische waarheidswaarde $(1, 1)$ (d.i. een totale onzekerheid) immers altijd aanleiding geven tot de beslissing $\{T, F\}$.

In de context van stringvergelijking is aangetoond dat de conditionele necessiteit moet worden bepaald met een kwantorgebaseerde aanpak. Dit is een gevolg van het niet-uniek zijn van de één-op-één afbeelding die wordt gebruikt. Een dergelijke aanpak zou ook toegepast kunnen worden in de context van complexe objecten. Immers, de vereiste van een gedeelde objectruimte impliceert een vaste één-op-één afbeelding. Bovendien kan in een meer algemene context waarin de objectruimte niet per definitie wordt gedeeld door twee objecten, steeds een één-op-één afbeelding tussen eigenschappen worden voorzien. Een duidelijk nadeel van deze aanpak is dat het verschil in discriminerende kracht tussen eigenschappen wordt verwaarloosd. Het feit dat kwantorfuncties snel en eenvoudig kunnen worden vastgelegd, spreekt dan weer in het voordeel van deze strategie. We zullen het beste van beide werelden proberen te combineren in een aanpak die de eenvoud van bepaling bewaart, maar die een intelligent model biedt waarbij de conditionele necessiteit van een groep van eigenschappen niet enkel afhangt van het aantal eigenschappen. Om dit te doen, bestuderen we twee aspecten van de conditionele necessiteit.

Ten eerste kunnen we aantonen dat het volstaat om een drastische γ^F te beschouwen. Immers, gelet op het feit dat de uiteindelijke beslissing van coreferentie wordt genomen door een beslissingsmodel (Hoofdstuk 5), kunnen we,

gelet op Definitie 5.1, stellen dat:

$$\forall \widetilde{p} \in \mathcal{F}(\mathbb{B}) : \mu_{\widetilde{p}}(F) = 1 \Rightarrow \mathscr{B}(\widetilde{p}) = F. \tag{7.35}$$

ongeacht $\mathscr{B}$. Objectkoppels waarvoor er geen zekerheid op coreferentie bestaat worden bijgevolg gelijk behandeld, ongeacht de zekerheid die er bestaat voor het niet-coreferent zijn van die koppels. Dit wil inderdaad zeggen dat γ^F mag worden beperkt tot een drastische functie. Deze veronderstelling maken we dan ook in wat volgt.

Ten tweede duiden we op de volgende stelling.

Stelling 7.3 (Wet van ononderscheidbaarheid)
Voor een Boolese propositie p, een verzameling van actoren A en een koppel van vertrouwensmaten $\gamma^{T,F}$ geldt er dat als:

$$\exists Q \subset A : \gamma^T(Q) \notin \{0,1\} \vee \gamma^F(Q) \notin \{0,1\} \tag{7.36}$$

dan bestaan er minstens twee verschillende verzamelingen $P_{\pi,1}$ en $P_{\pi,2}$ waarvoor:

$$S_{\gamma^{T,F}}(P_{\pi,1}) = S_{\gamma^{T,F}}(P_{\pi,2}). \tag{7.37}$$

Bewijs. Veronderstel dat er een $Q \subset A$ zodat $\gamma^T(Q) \notin \{0,1\}$ en beschouw $P_{\pi,1}$ zodat:

$$\forall \mathcal{A}_i \in Q \quad : \quad \widetilde{p}_i = (1,0) \tag{7.38}$$
$$\forall \mathcal{A}_i \notin Q \quad : \quad \widetilde{p}_i = (0,1). \tag{7.39}$$
$$\tag{7.40}$$

Omwille van de monotoniteit van γ^T geldt er dan:

$$S_{\gamma^{T,F}}(P_{\pi,1}) = \gamma^T(Q). \tag{7.41}$$

Beschouw nu $P_{\pi,2}$ zodat:

$$\exists \mathcal{A}_j \in Q \quad : \quad \gamma^T(Q) \leq \big(1 - \mu_{\widetilde{p}_j}(F)\big) < 1 \tag{7.42}$$
$$\forall \mathcal{A}_i \in Q \quad : \quad (\mathcal{A}_i \neq \mathcal{A}_j) \Rightarrow (\widetilde{p}_i = (1,0)) \tag{7.43}$$
$$\forall \mathcal{A}_i \notin Q \quad : \quad \widetilde{p}_i = (0,1). \tag{7.44}$$
$$\tag{7.45}$$

Dan geldt er enerzijds $P_{\pi,1} \neq P_{\pi,2}$ en anderzijds:

$$S_{\gamma^{T,F}}(P_{\pi,1}) = S_{\gamma^{T,F}}(P_{\pi,2}) = \gamma^T(Q). \tag{7.46}$$

Een analoge constructie kan worden gemaakt voor γ^F. □

De wet van ononderscheidbaarheid stelt dat wanneer conditionele necessiteit niet binair is, dan bestaan er steeds minstens twee verschillende verzamelingen van possibilistische waarheidswaarden die hetzelfde resultaat geven na combinatie met $S_{\gamma^{T,F}}$. Dit betekent dat de beslissing na toepassing van een beslissingsmodel $\mathscr{B}$ of $\mathscr{D}$ dezelfde is voor beide verzamelingen van possibilistische waarheidswaarden.

Voorbeeld 7.1
Beschouw p en $A = \{\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3\}$ en γ^T zodat er geldt dat:

$$\forall Q \subseteq A : \gamma^T(Q) = \begin{cases} 0.5 & \textbf{\textit{als}} \quad Q = \{\mathcal{A}_1, \mathcal{A}_2\} \\ 1 & \textbf{\textit{als}} \quad Q = A \\ 0 & \textbf{\textit{anders}}. \end{cases} \tag{7.47}$$

Beschouw dan de volgende verzamelingen van possibilistische waarheidswaarden:

$$P_{\pi,1} = \{(1,0),(1,0.1),(0,1)\} \tag{7.48}$$
$$P_{\pi,2} = \{(1,0),(1,0.4),(0,1)\} \tag{7.49}$$

waarvoor er inderdaad geldt dat:

$$S_{\gamma^{T,F}}(P_{\pi,1}) = S_{\gamma^{T,F}}(P_{\pi,2}) = (1,0.5). \tag{7.50}$$

Bijgevolg geldt er voor een willekeurig beslissingsmodel $\mathscr{B}$ dat:

$$\mathscr{B}\left(S_{\gamma^{T,F}}(P_{\pi,1})\right) = \mathscr{B}\left(S_{\gamma^{T,F}}(P_{\pi,2})\right). \tag{7.51}$$

Dit verschijnsel wordt verklaard door het feit dat conditionele necessiteit functioneert als een bovengrens voor zekerheid. Als we Voorbeeld 7.1 vertalen naar een context van coreferentie, dan wil dit zeggen dat er twee objectkoppels (o_1, o_2) en (o_3, o_4) bestaan waarvoor evaluator E_{U_2} een grotere zekerheid op coreferentie aanbiedt voor koppel (o_1, o_2), maar waarvoor de totale afgeleide zekerheid op coreferentie voor beide koppels dezelfde is. Om dergelijke situaties te vermijden moet er gelden dat:

$$\forall Q \subseteq A : \gamma^T(Q) \in \{0,1\}. \tag{7.52}$$

Merk op dat dezelfde voorwaarde reeds van toepassing is voor γ^F aangezien γ^F drastisch is. Deze vaststellingen over conditionele necessiteit leiden ons naar een aanpak waarbij γ^F volledig bepaald is door γ^T en waarbij γ^T een binaire functie is. Een bijzonder gevolg hiervan is dat we voor een willekeurige groep van actoren Q geen numerieke inschatting moeten maken van $\gamma^T(Q)$. Het volstaat om aan te geven of Q voldoende zekerheid biedt om te stellen dat $p = T$, aangenomen dat actoren in Q postuleren dat $p = T$. Dit vereenvoudigt de constructie van γ^T aanzienlijk. Ondanks deze eenvoud zal in Sectie 7.4 worden aangetoond dat de verkregen conditionele necessiteit een behoorlijk model biedt voor coreferentie van complexe objecten.

Alle elementen zijn nu voorhanden om onze aanpak voor het bepalen van conditionele necessiteit te definiëren. We veronderstellen zoals steeds een datacollectie $D \subset O^2$ waarbij C de verzameling van alle coreferente koppels in D voorstelt. We noteren $\widehat{C}$ als de benadering voor de verzameling van coreferente koppels door E_O. Dit is, in navolging van de notaties in Sectie 6.7.2, de verzameling van alle koppels waarvoor een necessiteit voor coreferentie groter

dan 0 bestaat. Door gebruik te maken van deze notatie kunnen *zuiverheid* en *compleetheid* worden gedefinieerd als:

$$\text{zuiverheid} = \frac{|C \cap \widehat{C}|}{|\widehat{C}|} \tag{7.53}$$

$$\text{compleetheid} = \frac{|C \cap \widehat{C}|}{|C|}. \tag{7.54}$$

We zullen *zuiverheid* afkorten tot *zuiv* en *compleetheid* tot *comp* om de notaties compact en overzichtelijk te houden. In onze aanpak veronderstellen we dat deelevaluatoren E_{U_i} een hoge lokale *compleetheid* hebben. Dit betekent dat wanneer coreferente complexe objecten worden geprojecteerd naar U_i, de kans hoog is dat de resulterende deelobjecten een necessiteit op coreferentie groter dan 0 toegekend krijgen door E_{U_i} en dus als coreferent worden gezien. In Hoofdstuk 6 is onderzocht hoe dit kan voor strings. Gelet op deze veronderstelling van hoge lokale *compleetheid* moet de bepaling van γ^T, verzamelingen van eigenschappen vinden die tot een hoge *zuiverheid* leiden. Dit kan worden ingezien als volgt. Een koppel (o_1, o_2) wordt een kandidaat volgens E_{U_i} genoemd als:

$$\mathscr{B}\left(E_{U_i}\left(\text{proj}_i(o_1), \text{proj}_i(o_2)\right)\right) = T \tag{7.55}$$

voor een maximaal tweewaardig beslissingsmodel $\mathscr{B}$. Een koppel (o_1, o_2) is dus een kandidaat volgens E_{U_i} als er zekerheid op coreferentie bestaat voor de i^{de} projectie van beide objecten. Als de deelevaluatoren een hoge lokale *compleetheid* hebben, dan wil dit zeggen dat coreferente koppels kandidaat zijn volgens de meeste E_{U_i}. Een typisch gevolg hiervan is dat heel wat kandidaten geen coreferente koppels zijn, hetgeen een lage *zuiverheid* kan impliceren. Dit probleem kunnen we oplossen door te eisen dat een koppel (o_1, o_2) kandidaat moet zijn volgens meerdere deelevaluatoren, waardoor de *zuiverheid* wordt verhoogd. Wanneer de *zuiverheid* bereikt door een combinatie van deelevaluatoren hoog genoeg is, zullen we stellen dat deze evaluatoren een groep van actoren impliceren waarvoor de conditionele necessiteit 1 is. Een probleem treedt op wanneer een eigenschap voor vele entiteiten niet meetbaar is. Niet-meetbare eigenschappen zullen aanleiding geven tot een complete onzekerheid en dus een necessiteit gelijk aan 0 opleveren (Hoofdstuk 2). Wanneer een eigenschap wordt aangetroffen die voor vele objecten niet meetbaar is, dan kan dit de veronderstelling van hoge *compleetheid* verstoren en moet hiermee rekening worden gehouden.

Voor elke mogelijke verzameling van actoren Q zullen we dus stellen dat $\gamma^T(Q) = 1$ als en alleen als de benaderde *zuiverheid* op basis van Q hoog genoeg is. Hiervoor gebruiken we opnieuw de notatie $\widehat{C}_i$ uit vorige sectie, zodat (o_1, o_2) kandidaat is volgens E_{U_i} als:

$$(o_1, o_2) \in \widehat{C}_i. \tag{7.56}$$

Beschouw de indexverzameling $I_O = \{1, ..., n\}$. Voor een willekeurige verzame-

ling $J \subseteq I_O$ noteren we:

$$\widehat{C}_J = \{(o_1, o_2) | (o_1, o_2) \in D \wedge \forall j \in J : \mathscr{B}(E_{U_j}(\text{proj}_j(o_1), \text{proj}_j(o_2))) = T\} \tag{7.57}$$

met $\mathscr{B}$ een maximaal tweewaardig beslissingsmodel. Een benadering van de *zuiverheid* voor een combinatie van actoren Q wordt dan verkregen als:

$$\widehat{\text{zuiv}}_Q = \left(\frac{1}{|\widehat{C}_J|}\right) \min_{j \notin J} |\widehat{C}_j \cap \widehat{C}_J| \tag{7.58}$$

waarbij J de indexen bevat van de actoren uit Q. Dit betekent dat voor een groep van actoren Q de *zuiverheid* wordt benaderd als de kleinste overeenkomst tussen de kandidaatverzameling van Q en de kandidaatverzameling van een actor die niet in Q zit. Onze aanpak bestaat erin op zoek te gaan naar de laagste k^* zodat er minstens één Q bestaat waarvoor $|Q| = k^*$ en zodat de benaderde *zuiverheid* van Q voldoende hoog is. Dit betekent dat er moet gelden:

$$k^* = \min \left\{ k | \exists Q \subseteq A : |Q| = k \wedge \widehat{\text{zuiv}}_Q \geq \tau_{\text{zuiv}} \right\}. \tag{7.59}$$

Anders gezegd, de benaderde *zuiverheid* $\widehat{\text{zuiv}}_Q$ mag niet kleiner zijn dan een vooraf bepaalde drempelwaarde. Eens we deze k^* gevonden hebben, zullen we groepen van k^* actoren gaan selecteren om de conditionele necessiteit op te baseren. Hiervoor construeren we een verzameling Q^* die verzamelingen van k^* actoren bevat en zodat er geldt dat:

$$\forall Q_i^* \in Q^* : \frac{\widehat{\text{zuiv}}_{Q_i^*}}{\max_{Q \subseteq A, |Q| = k^*} \widehat{\text{zuiv}}_Q} \leq \tau_{\text{sel}} \tag{7.60}$$

en

$$\frac{1}{|D|} \left| \left\{ (o_1, o_2) \middle| \bigvee_{Q_i^* \in Q^*} \bigwedge_{j \in J_i^*} \left(\text{proj}_j(o_1) \neq \perp \wedge \text{proj}_j(o_2) \neq \perp \right) \right\} \right| \leq \tau_{\perp}. \tag{7.61}$$

waarbij J_i^* de indexen van actoren in Q_i^* voorstelt. We herinneren eraan dat $\perp$ een (deel)object is dat het resultaat is van een meting van een niet-meetbare entiteit (of een eigenschap ervan). Anders gezegd, $\perp$ drukt uit dat bepaalde gegevens ontbreken. Voorwaarde (7.60) drukt uit dat combinaties van k^* actoren worden geselecteerd, waarvoor de benaderde *zuiverheid* voldoende dicht bij de maximale *zuiverheid* voor k^* actoren ligt. Voorwaarde (7.61) drukt uit dat combinaties van k^* actoren worden geselecteerd, zodat de eigenschappen waarop deze actoren hun kennis baseren voldoende meetbaar zijn binnen de datacollectie D. De conditionele necessiteit kan dan worden bepaald als:

$$\forall Q \subseteq A : \gamma^T(Q) = \begin{cases} 1 & \textbf{als} \quad \exists Q_i^* \in Q^* : Q_i^* \subseteq Q \wedge |Q| > \max(k^*, |Q^*|) \\ 0 & \textbf{anders} \end{cases} \tag{7.62}$$

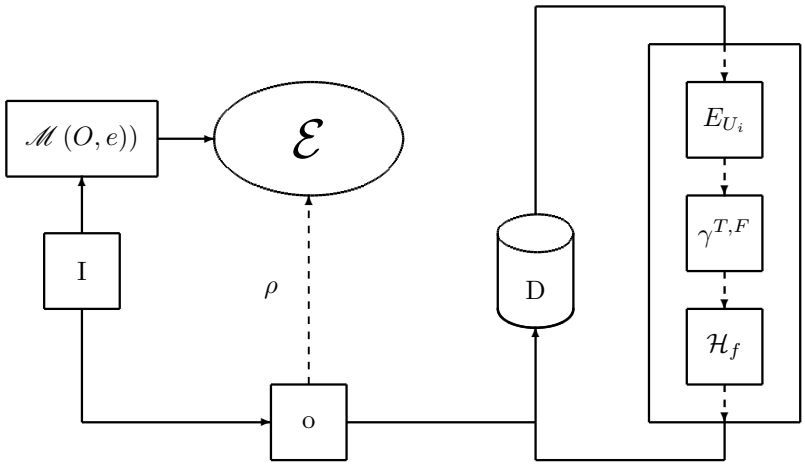
De bijkomende restrictie op de kardinaliteit van Q wordt toegevoegd om te verzekeren dat $\gamma^T(Q) = 1$ als en alleen als $|Q|$ voldoende groot is. Hiermee worden situaties vermeden waarbij $|Q| = 1$ aanleiding geeft tot $\gamma^T(Q) = 1$, hetgeen zich zou kunnen voordoen als een actor een zeer hoge benaderde *zuiverheid* heeft[3]. Experimenten tonen echter aan dat actoren op zich nooit voldoende conditionele necessiteit kunnen bieden om een goed resultaat te verkrijgen.

In wat volgt zal blijken dat de verkregen vertrouwensmaten meestal zeer eenvoudig zijn, maar toch een goed model voor coreferentie vormen. We herhalen hier dat de verkregen conditionele necessiteit kan worden aangepast door gebruik van een groeperingsfunctie λ. Na bepaling van γ^T kunnen we namelijk de functie γ^T_λ construeren zodat er voor elke $Q = \{\mathcal{A}_1, ..., \mathcal{A}_k\}$ geldt dat:

$$\gamma^T_\lambda(Q) = \min\left(\gamma^T(Q), \lambda(\{\text{lab}(U_1), ..., \text{lab}(U_k)\})\right). \tag{7.63}$$

7.3.5 Beslissingsmodellen

We hebben nu beschreven hoe een evaluator E_O kan worden geconstrueerd voor een datacollectie $D \subset O^2$. Zowel bij de constructie van deelevaluatoren (Hoofdstuk 6) als bij de constructie van conditionele necessiteit zijn we steeds uitgegaan van een maximaal beslissingsmodel (Hoofdstuk 5).



Figuur 7.2: Schematisch overzicht van het possibilistisch raamwerk

Uit de eigenschappen van beslissingsmodellen die zijn bestudeerd in Hoofdstuk 5 volgt dat elk ander beslissingsmodel aanleiding geeft tot een deelverzameling van coreferente koppels ten opzichte van de verzameling verkregen voor het maximale beslissingsmodel. Door het gebruik van de voorgestelde algoritmen voor het genereren van een partitie over O (Hoofdstuk 5) kan een consistent

[3]Hierbij veronderstellen we impliciet dat een complex universum $O = U_1 \times ... \times U_n$ uit minstens twee deeluniversa bestaat (d.i., $n \geq 2$). Dit heeft geen invloed op de algemeenheid vermits in het geval van $n = 1$ geen conditionele necessiteit moet worden berekend.

model voor coreferentie worden opgebouwd. We hebben in dit hoofdstuk geen aandacht besteed aan driewaardige beslissingsmodellen omdat in Hoofdstuk 5 is aangetoond hoe een driewaardig beslissingsmodel kan worden opgebouwd vertrekkende van een tweewaardig beslissingsmodel. Het possibilistische raamwerk voor coreferentiebepaling wordt schematisch samengevat in Figuur 7.2. Deze figuur toont objecten o in een databank D. Objecten verwijzen via ρ naar entiteiten in $\mathcal{E}$. Deze verwijzing is het resultaat van een meetproces $\mathscr{M}$. Het zoeken naar coreferente objecten gebeurt door eerst deelevaluatoren E_{U_i} te kiezen en de parameters van deze deelevaluatoren te bepalen. Vervolgens wordt de conditionele necessiteit $\gamma^{T,F}$ bepaald. Als laatste stap wordt een partitie van objecten bekomen waarbij coreferente objecten samen in een partitieklasse zitten.

7.4 Experimenten

In deze sectie zullen we een aantal experimentele resultaten presenteren waarin onze aanpak wordt vergeleken met enkele methoden uit de literatuur. Hierbij wordt grotendeels de testopstelling uit Sectie 6.8 hernomen, maar nu in de context van complexe objecten in plaats van strings. De gebruikte datacollecties worden opgesomd in Tabel 7.1.

naam	$\|D\|$	$\|C\|$	**bron**
census	176008	327	Winkler
cora	837865	17184	[117]
hotels	131075	229	-
restaurant	176423	112	[102]

Tabel 7.1: De gebruikte datacollecties

Tabel 7.2 toont detailinformatie omtrent de gebruikte datacollecties. Voor elke datacollectie wordt voor elke eigenschap het aantal verschillende objecten na projectie getoond. Voor de 'census' datacollectie zijn er bijvoorbeeld 606 verschillende deelobjecten die de eigenschap 'firstname' beschrijven. Daarnaast wordt het percentage ontbrekende waarden voor elke eigenschap gegeven. Vermits alle deelobjecten strings zijn, kan ook voor elke eigenschap het gemiddeld aantal deelstrings na splitsing worden berekend. Drie van deze datacollecties worden frequent vermeld in de literatuur die handelt over coreferentie. De 'hotels' datacollectie is een nieuwe datacollectie die is opgebouwd in het kader van deze thesis. Hiervoor zijn twee websites geraadpleegd[4] waarop hotels geboekt kunnen worden. Voor drie datacollecties ('census', 'hotels' en 'restaurant') zijn twee verschillende collecties van objecten voorhanden. Voor de andere ('cora') is één collectie voorhanden en moet aan zelfvergelijking worden gedaan. Voor enkele steekproeven uit deze datacollecties verwijzen we naar Bijlage E. We

[4]http://www.booking.com en http://www.boekuwhotel.be

naam	eigenschap	waarden	$\%\perp$	gem. $\lvert\mathscr{S}_{\mathscr{G}}(s)\rvert$
census	firstname	606	0.14	1
census	middlename	24	0.20	1
census	lastname	323	0.00	1
census	street	39	0.00	1
census	housenumber	170	0.00	1
cora	title	292	0.00	8
cora	author	486	0.00	7
cora	pages	305	0.33	3
cora	venue	513	0.10	8
cora	year	70	0.12	1
hotels	name	651	0.00	3
hotels	address	776	0.00	4
hotels	stars	8	0.01	1
hotels	city	14	0.00	1
restaurant	name	776	0.00	2
restaurant	street	772	0.00	4
restaurant	city	49	0.00	2
restaurant	type	84	0.00	1

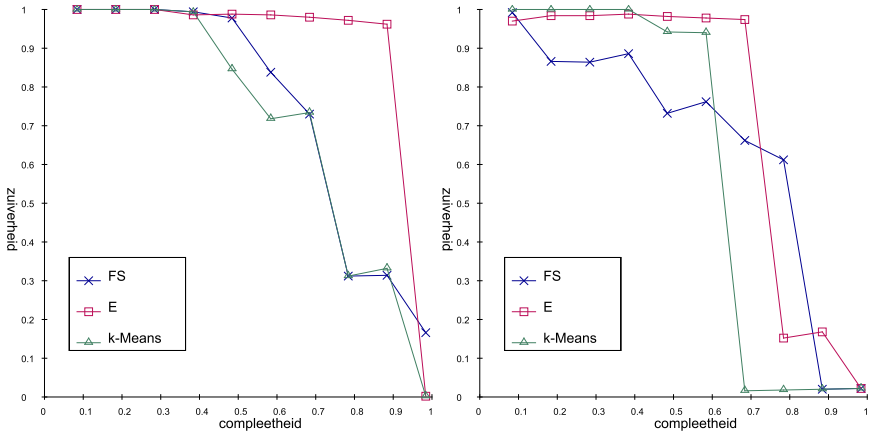
Tabel 7.2: Details van de datacollecties

construeren voor onze aanpak telkens een evaluator E_O, zodat een *zuiverheid-compleetheidscurve* kan worden opgebouwd, zoals beschreven in Sectie 6.8. Zo kan ook de maximale f-waarde worden berekend overheen de verschillende niveaus van *compleetheid*.

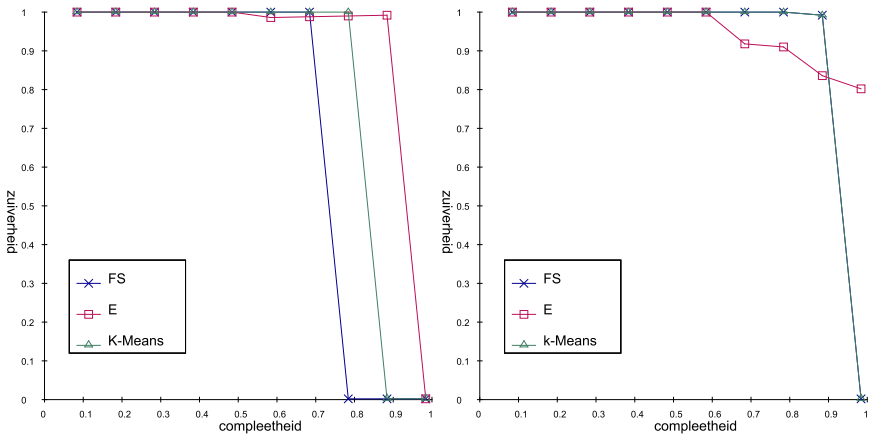
Voor onze possibilistische aanpak kiezen we als drempelwaarden $\tau_{\text{sel}} = 0.80$ en $\tau_{\perp} = 0.01$. De k-snede voor relaties gebruikt door semantische evaluatoren wordt bepaald voor $k = 5$. Deze waarden zijn empirisch vastgesteld en het is onze bedoeling om deze parameterwaarden vast te kiezen, ongeacht het precieze probleem dat wordt bestudeerd. Om onze aanpak te kunnen vergelijken met bestaande methoden, maken we gebruik van de Java-bibliotheek FEBRL [118], waarin verschillende technieken voor coreferentiebepaling zijn geïmplementeerd. We rapporteren in onze vergelijkende studie de resultaten van de Fellegi-Sunter aanpak (FS) met optimale drempelwaarden en *k-means* clustering met (k=2).

Figuren 7.3 en 7.4 tonen de *zuiverheid-compleetheidscurven* voor *k-means* en het FS-model vergeleken met de possibilistische aanpak (E). Een interessant verschil is merkbaar voor datacollectie 'restaurant' (Figuur 7.4), waar de voorkeur van onze aanpak voor *compleetheid* over *zuiverheid* duidelijk te merken is. Onze methode heeft namelijk een veel hogere *zuiverheid* bij *compleetheid* gelijk aan 1. Voor de andere datacollecties is een gelijkaardige trend merkbaar, zij het in mindere mate.

Tabel 7.3 toont een vergelijking van de maximale f-waarden, berekend uit de verschillende *zuiverheid-compleetheidscurven*. Voor 'restaurant' wordt de



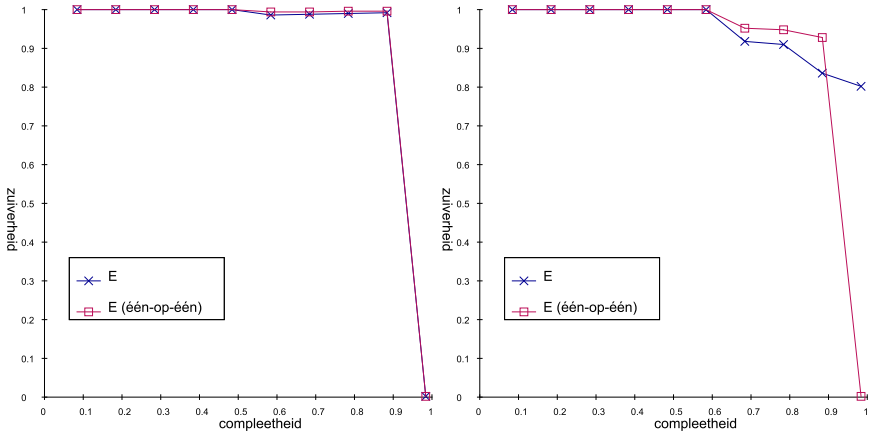
Figuur 7.3: *Zuiverheid* vs. *compleetheid* voor 'census' (links) en 'cora' (rechts)



Figuur 7.4: *Zuiverheid* vs. *compleetheid* voor 'hotels' (links) en 'restaurant' (rechts)

naam	E_O	k-means	FS
census	**0.929**	0.717	0.714
cora	**0.813**	0.732	0.693
hotels	**0.943**	0.889	0.824
restaurant	0.889	**0.943**	**0.943**

Tabel 7.3: Maximale f-waarden

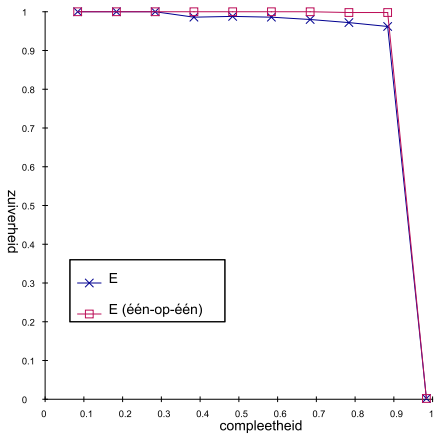


Figuur 7.5: De invloed van randvoorwaarden: 'hotels' (links) en 'restaurant' (rechts)

voorkeur voor *compleetheid* afgestraft met een lagere maximale f-waarde. Voor de andere datacollecties toont de possibilistische aanpak een hogere maximale f-waarde.

In Hoofdstuk 5 is besproken hoe inconsistenties van een evaluator kunnen worden aangepakt en hoe randvoorwaarden kunnen worden opgelegd. In Hoofdstuk 6 is aangetoond dat het opleggen van randvoorwaarden in het geval van strings, een positief effect heeft op de kwaliteit van onze aanpak. We willen dit experiment herhalen in de context van complexe objecten. Voor drie van de vier datacollecties geldt er dat twee afzonderlijke collecties worden vergeleken. Voor deze datacollecties kunnen we de randvoorwaarde opleggen dat elk object uit de ene collectie maximaal met één object uit de andere collectie coreferent kan zijn. Wanneer we deze randvoorwaarde opleggen met de methode uit Hoofdstuk 5, dan verkrijgen we de resultaten getoond in Figuren 7.5 en 7.6.

Het opleggen van de randvoorwaarden heeft tot gevolg dat de mogelijkheid op coreferentie voor bepaalde koppels wordt weggenomen. Bij datacollecties 'hotels' en 'census' heeft dit een positieve invloed op de *zuiverheid*. Bij datacollectie 'restaurant' valt het op dat er een aantal coreferente koppels zijn die door E_O een bepaalde mogelijkheid op coreferentie worden toegekend, maar waarbij die mogelijkheid door de randvoorwaarden wordt weggenomen. Het afdwingen van de randvoorwaarde zorgt dus in dit geval voor een hogere *zuiverheid* ten koste van een lagere *compleetheid*.



Figuur 7.6: De invloed van randvoorwaarden: 'census'

In de datacollectie 'cora' kan de randvoorwaarde niet worden opgelegd. Voor deze datacollectie zullen we onderzoeken wat de invloed van herstel van transitiviteit is. Na constructie van de evaluator E_O beschouwen we een beslissingsmodel $\mathscr{B}$ met drempelwaarde $z = 0$ (Hoofdstuk 5). Dit betekent dat twee objecten als coreferent worden beschouwd, als er een zekerheid voor waar bestaat die groter is dan nul. We zullen hier de functie

$$\mathcal{H}_{f_{S_{\gamma^T,F}}} \tag{7.64}$$

gebruiken om het niet-transitief zijn van E_O te compenseren (Hoofdstuk 5). Op basis van de verkregen partitieklassen herberekenen we vervolgens de *zuiverheid*. Dit betekent enerzijds dat twee objecten o_1 en o_2 uit dezelfde partitieklasse, waarvoor er geldt dat:

$$\mathscr{B}(E_O(o_1, o_2)) = F \tag{7.65}$$

als coreferent worden gezien. Anderzijds betekent dit ook dat twee objecten o_1 en o_2 uit verschillende partitieklassen, waarvoor er geldt dat:

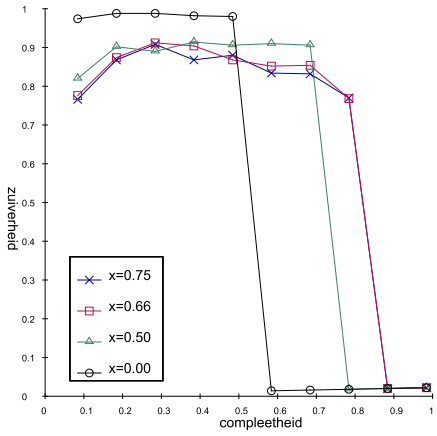
$$\mathscr{B}(E_O(o_1, o_2)) = T \tag{7.66}$$

niet als coreferent worden gezien. Wanneer we daarmee rekening houden, verkrijgen we een ander verloop van *zuiverheid*. Voor de constructie van conditionele necessiteit maken we gebruik van de functie:

$$q_x(n, m) = \begin{cases} 1 & \textbf{als} \quad \frac{n}{m} \leq x \\ 0 & \textbf{als} \quad \frac{n}{m} > x \end{cases} \tag{7.67}$$

met $x \in [0, 1]$ zodat:

$$\gamma^T(Q) = q_x(|Q|, |A|). \tag{7.68}$$



Figuur 7.7: De invloed van $\mathcal{H}_f$: 'cora'

Figuur 7.7 toont het verloop van *zuiverheid* over standaardwaarden van *compleetheid* voor vier verschillende waarden van x. Merk op dat voor $x = 0$ de functie $\mathcal{H}_{f_{\hat{\wedge}}}$ wordt verkregen. We merken duidelijk dat voor lage waarden van x enkel aandacht wordt besteed aan *zuiverheid*. Naarmate x groter wordt, worden de voorwaarden voor partitieklassen soepeler, zodat de *compleetheid* belangrijker wordt. Dit kunnen we inzien doordat voor $x = 0.66$ en $x = 0.75$ de *zuiverheid* bij *compleetheid* 0.8 aanzienlijk toeneemt. Een gevolg is echter dat de *zuiverheid* voor lagere waarden van de *compleetheid* daalt.

Wanneer de possibilistische aanpak afzonderlijk wordt beschouwd, kan een interessante vaststelling worden gedaan. Voor 8 eigenschappen wordt een twee-niveau evaluator toegekend en wordt een bijhorende kwantorfunctie bepaald. De resulterende parameters voor deze kwantorfuncties liggen bijzonder dicht bij elkaar. Rekening houdend met het feit dat het model van een twee-niveau evaluator voor het vergelijken van strings kan worden gezien als:

"Twee strings zijn coreferent als de meeste deelstrings coreferent zijn",

waarbij de kwantorfunctie een model vormt voor "de meeste", wil dit zeggen dat de interpretatie van "de meeste" sterk gelijkaardig is overheen verschillende problemen. Tabel 7.2 toont dat het gemiddeld aantal deelstrings voor de eigenschappen sterk varieert, hetgeen de conclusie ondersteunt dat kwantificatie in de context van coreferentie een generiek gegeven is, eerder dan probleemspecifiek.

7.5 Conclusie

In dit hoofdstuk is een possibilistische aanpak voor het coreferentieprobleem in de context van complexe objecten geïntroduceerd. Onze oplossing vertrekt

van de vaststelling dat de coreferentiële propositie van twee complexe objecten ($p_{(o_1,o_2)}$) kan worden ontbonden in n coreferentiële deelproposities. We kunnen dan n actoren beschouwen die uitspraken doen over $p_{(o_1,o_2)}$, waarbij elke actor zich baseert op één deelpropositie. De constructie van een evaluator E_O vertrekt van een toekenning van deelevaluatoren aan de deeluniversa waaruit een complex universum O is opgebouwd. Elke deelevaluator zal de kennis over exact één deelpropositie genereren. In een tweede stap worden deze evaluatoren aangepast aan de gegeven datacollectie. Dit deelprobleem is in de context van strings besproken in vorige hoofdstukken. In een derde en laatste stap wordt het vertrouwen in actoren in kaart gebracht door de constructie van twee vertrouwensmaten γ^T en γ^F. Het bipolaire karakter van deze twee vertrouwensmaten is volledig compatibel met driewaardige beslissingsmodellen. Het is aangetoond dat de combinatie van een evaluator E_O (kennisgeneratie) met een beslissingsmodel $\mathscr{B}$ (kennisinterpretatie) tot gevolg heeft dat beide vertrouwensmaten vereenvoudigd kunnen worden tot drastische en binaire vertrouwensmaten. Als gevolg hiervan verkrijgen we een bijzonder eenvoudig mechanisme voor het bepalen van conditionele necessiteit. Experimenten tonen aan dat de verkregen evaluator E_O beter doet dan bestaande methoden uit de literatuur. De voordelen van onze aanpak zijn een eenvoudige bepaling van conditionele necessiteit die een minimum aan handmatige tussenkomst vereist, een elegante behandeling van niet-meetbaarheid en een toepasbaarheid op uiteenlopende problemen.

Hoofdstuk 8

Coreferentie van teksten

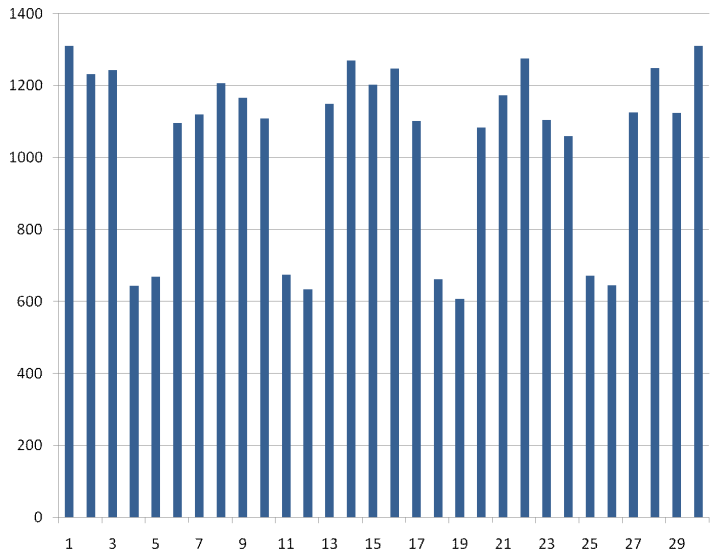
8.1 Inleiding

In voorgaande hoofdstukken is een gedetailleerde uiteenzetting gegeven over het coreferentieprobleem. Hierbij is verondersteld dat de beschrijving van een entiteit het resultaat is van een meting van vooraf bepaalde eigenschappen. De ruimte waarin deze eigenschappen worden beschreven is steeds vooraf gekend. Zoals besproken in Hoofdstuk 2 kunnen vaak bijkomende structurele beperkingen worden opgelegd, bijvoorbeeld door gebruik van een groeperingsfunctie λ. In dit hoofdstuk bespreken we een heel andere context van coreferentie, namelijk een context waarin elk object een tekstuele beschrijving is van een entiteit.

In Hoofdstuk 6 begint de studie van evaluatoren voor strings met de belangrijke veronderstelling dat de strings in kwestie een antwoord formuleren op de typevraag "Wat is eigenschap a van entiteit e?", zodat een complex object ontstaat uit een welomlijnde verzameling van deelobjecten die eigenschappen beschrijven. De typevraag wordt hier vervangen door de typevraag "Hoe zou u e beschrijven?", waarbij geen veronderstelling wordt gemaakt over de eigenschappen van e die deel moeten uitmaken van o. Het resultaat van een dergelijke vraag is een vrije, tekstuele beschrijving van een entiteit. Waar bij de eerste typevraag het antwoord relatief éénduidig bepaald is, op enkele causaliteiten van onzekerheid na, is dit voor de tweede typevraag niet meer het geval. We stellen ons hier de vraag of het voor dergelijke beschrijvende antwoorden eveneens mogelijk is om te bepalen welke beschrijvingen naar eenzelfde entiteit refereren. Dit probleem staat in de literatuur doorgaans bekend als tekstclustering of documentclustering. Hoewel strikt genomen de term 'document' algemener is dan de term 'tekst', zullen we in wat volgt beide termen als onderling equivalent beschouwen.

Aangezien de toepassingen van coreferentie voor complexe objecten al uitvoerig zijn besproken, willen we deze oefening herhalen voor het geval van tekstuele beschrijvingen. Een eerste argument dat de noodzaak aan goede

tekstclustering duidelijk maakt, is de dagelijkse hoeveelheid nieuwe informatie die op het web beschikbaar wordt gemaakt. Beschouw bijvoorbeeld de talrijke nieuwssites op het WWW, zoals er ook verschillende in het Nederlands beschikbaar zijn. Het gebruik van RSS-berichten (Really Simple Syndication) door dergelijke sites laat ons toe een inschatting te maken van het dagelijks aantal gepubliceerde nieuwsartikels. Figuur 8.1 toont de evolutie over dertig dagen van het dagelijks aantal nieuwe RSS-berichten dat gezamenlijk beschikbaar wordt gemaakt door tien verschillende Nederlandstalige nieuwssites.



Figuur 8.1: Evolutie van de hoeveelheid nieuwe RSS-berichten van tien nieuwssites

Aangezien geen enkele nieuwssite voor elk nieuw verschenen artikel een RSS-bericht stuurt, vormen de cijfers uit Figuur 8.1 een ondergrens voor het dagelijks aantal nieuwe artikels dat op de beschouwde Nederlandstalige nieuwssites verschijnt. De conclusie van deze metingen is duidelijk: het dagelijks aantal nieuwe artikels voor Nederlandstaligen valt onmogelijk handmatig te verwerken. Uiteraard kan een criticaster van tekstclustering argumenteren dat de informatie die wordt gedeeld door verschillende nieuwssites, aanzienlijk is. Hierdoor kan een actor zich beperken tot een minimum aan informatiebronnen. Een dergelijke redenering is slechts gedeeltelijk correct. Enerzijds wordt in deze thesis erkend dat het gebruik van meerdere informatiebronnen in parallel leidt tot herhaalde data. Sterker nog, deze vaststelling speelt een belangrijke rol bij de inschatting van het aantal clusters (Sectie 8.5). Anderzijds verwerpen we de redenering dat actoren zich daarom kunnen beperken tot een minimum aantal nieuwsbronnen. Om dit argument kracht bij te zetten, verwijzen we naar een studie waarin met een testpubliek wordt aangetoond dat mensen beter in staat zijn feitelijke gegevens te verzamelen wanneer een collectie van coreferente tek-

sten samengevat wordt tot één tekst [119]. Vanzelfsprekend is het automatisch genereren van dergelijke samenvattingen enkel mogelijk als de kwaliteit van de gevormde tekstclusters voldoende hoog is. De noodzaak aan accurate methoden voor het identificeren van coreferente teksten wordt hierdoor benadrukt.

In wat volgt zal het concrete geval van nieuwswebsites verder worden onderzocht. Hiervoor zijn 550 Nederlandstalige documenten verzameld, gespreid over een tijdsperiode van twee maanden en ingeladen vanuit een tiental Belgische en Nederlandse websites. Het feit dat ook websites uit Nederland gebruikt zijn, vermindert de kans dat overheen verschillende documenten, grote stukken tekst worden herbruikt. Aan elk document is manueel een onderwerp toegekend door één persoon. Twee documenten worden eenzelfde onderwerp toegekend als één van beide documenten expliciet verwijst naar het onderwerp van het andere document. In totaal zijn er op die manier 133 onderwerpen vertegenwoordigd in de verzameling van documenten. Hierbij zijn enkele moeilijkheden geïntroduceerd om de datacollectie representatief te maken. Zo zijn er documenten over vier verschillende vliegtuigongelukken verzameld, waarbij het automatisch onderscheid maken tussen deze ongelukken een uitdaging vormt. Ook zijn er documenten aanwezig over gebeurtenissen die zich op eenzelfde plaats afspelen, bijvoorbeeld drie verschillende onderwerpen die zich op Schiphol afspelen. Er zijn tevens heel wat documenten verzameld over randgebeurtenissen, waardoor onderwerpen niet eenvoudig af te lijnen zijn. De moeilijkheid van deze datacollectie kan worden ingeschat door de resultaten van bestaande methoden voor tekstclustering te bestuderen in Sectie 8.7. We introduceren deze datacollectie uitzonderlijk aan het begin van dit hoofdstuk om de voorbeelden doorheen dit hoofdstuk beter te kaderen. Voor enkele voorbeelden van tekstdocumenten uit de gebruikte datacollectie verwijzen we naar Bijlage E.

Naar goede gewoonte geven we in Sectie 8.2 eerst een overzicht van de belangrijkste literatuur over bestaande oplossingen voor het probleem van coreferente documenten. Vertrekkende van de tekortkomingen van deze oplossingen wordt een nieuwe aanpak voorgesteld. In Sectie 8.3 introduceren we een nieuw model om tekst voor te stellen. In dit model wordt een wiskundige abstractie gemaakt van het i.Know-tekstmodel[1], een gepatenteerde technologie die enkele veelbelovende voordelen biedt. We zullen op basis van het nieuwe model komen tot een evaluator voor documenten in Sectie 8.4. Het zal blijken dat een dergelijke evaluator context-afhankelijk werkt. Dit betekent dat twee documenten enkel vergeleken kunnen worden als de context waarin deze documenten voorkomen, gekend is. We zullen aantonen dat het daarom efficiënter is om rechtstreeks een partitie te maken van een verzameling van documenten. In Sectie 8.5 zal daarom een nieuwe methode worden gegeven om het aantal partitieklassen (d.i. het aantal clusters) te bepalen. Het bepalen van het aantal clusters is volgens ons nog onvoldoende onderzocht in het kader van tekstclustering en is van cruciaal belang om een goede clustermethode te kunnen ontwikkelen. In Sectie 8.6 wordt een mechanisme beschreven dat collecties

[1] http://www.iknow.be

van coreferente documenten zoekt in een gegeven verzameling van documenten, hetgeen leidt tot een nieuw mechanisme voor het clusteren van tekstuele beschrijvingen van entiteiten. In Sectie 8.7 vergelijken we onze aanpak met geavanceerde technieken uit de literatuur. Sectie 8.8 geeft een overzicht van de belangrijkste bevindingen uit dit hoofdstuk.

8.2 Vectorruimtemodel

Het klassieke model om documenten te modelleren is het vectorruimtemodel, dat oorspronkelijk werd ontworpen voor indexatie van teksten [120]. Hierbij wordt uitgegaan van een verzameling van documenten $D = \{d_1, ..., d_n\}$. Voorts wordt verondersteld dat er een verzameling $\mathcal{G}$ van q-grammen bestaat. Een q-gram is hierbij een woordgroep van q woorden. Een document d_i wordt gemodelleerd als een vector $\mathbf{v}^i$ waarbij $\mathbf{v}^i_j$ de relevantie aangeeft van de j^{de} q-gram voor document d_i. De manier waarop relevantie wordt aangeduid, hangt af van de toepassing. In het meest triviale geval is $\mathbf{v}^i_j$ binair en geeft het aan of de j^{de} q-gram voorkomt in document d_i. In de meest voorkomende gevallen geldt er $\mathbf{v}^i_j \in [0, 1]$ en wordt er voor de j^{de} q-gram een graduele relevantie berekend, zoals het TFIDF gewicht [89] (Hoofdstuk 6). Aangezien de dimensie van vectoren constant wordt verondersteld, kan de verzameling D worden voorgesteld als een matrix:

$$\mathbf{M}_D = \left[\mathbf{v}^i_j\right]. \tag{8.1}$$

Een voordeel van dit model is dat de matrixvoorstelling toelaat om klassieke clustermethoden toe te passen op dit probleem. Voor een overzicht van de belangrijkste methoden voor het clusteren van teksten verwijzen we naar Bijlage A. In dit hoofdstuk stellen we ons de vraag of we niet beter kunnen doen dan deze klassieke clustermethoden en dit om verschillende redenen. Ten eerste heeft $\mathcal{G}$ een zeer sterk oplopende kardinaliteit in functie van stijgende n. De vectoren $\mathbf{v}^i$ kunnen bijgevolg een zeer hoge dimensie hebben. Een praktisch aanvaardbare oplossing wordt dan zowel vanuit het standpunt van correctheid van de resultaten als vanuit het standpunt van complexiteit onmogelijk zonder dimensiereductie. Technieken voor dimensiereductie brengen uiteraard een bijkomende algoritmische complexiteit met zich mee. Ten tweede is het q-gram model een weinig onderbouwde kunstgreep om tekstclustering te kunnen vertalen naar het vectorruimtemodel en op die manier klassieke methoden te kunnen toepassen. Het q-gram model is meestal enkel haalbaar voor $q = 1$ als men de algoritmische complexiteit wil beperken. Beide vernoemde problemen houden verband met een derde probleem, namelijk dat het vectorruimtemodel voor documenten de semantische achtergrond in documenten negeert. Dit is een tekortkoming die we in het nieuwe model wensen te voorkomen.

8.3 Relationeel documentmodel

Als alternatief voor het vectorruimtemodel wordt in deze thesis het relationeel model voor documenten voorgesteld. Om dit model te kunnen definiëren is het nodig het concept van binaire multirelaties te introduceren. Multirelaties laten zich echter niet eenduidig definiëren. Een binaire relatie[2] over het universum O is een deelverzameling van het Cartesisch product van O, namelijk O^2. Het begrip 'deelverzameling' laat zich eenduidig veralgemenen naar het raamwerk van multiverzamelingen. Een veralgemening van de verzameling O^2 naar het raamwerk van multiverzamelingen kan echter op twee manieren: als Cartesisch product van twee multiverzamelingen ($\mathcal{M}(O) \times \mathcal{M}(O)$) of als multiverzameling van het Cartesisch product ($\mathcal{M}(O \times O)$). De eerste definitie wordt aangehaald in [72] en impliceert een aantal bijzondere eigenschappen die nuttig zijn voor de definitie van similariteitsmaten voor multiverzamelingen. De tweede definitie lijkt ons de meest natuurlijke en het is deze definitie die hier zal worden gebruikt.

Definitie 8.1 (Binaire multirelatie)
Voor een universum O is een binaire multirelatie R over O gedefinieerd als een deelverzameling van de multiverzameling $\mathcal{M}(O \times O)$.

Deze definitie van het concept 'binaire multirelatie' kan worden gezien als een multiverzameling van koppels. Elk koppel in de relatie kan bijgevolg meerdere keren voorkomen en dit aantal wordt aangeduid door de multipliciteit van het koppel. Een binaire multirelatie R over O is symmetrisch als er geldt dat:

$$\forall (o_1, o_2) \in O^2 : \omega_R(o_1, o_2) = \omega_R(o_2, o_1). \tag{8.2}$$

Een binaire multirelatie R wordt k-transitief genoemd als de relatie R_k transitief is. Hierbij is R_k de k-snede van de multirelatie R, d.i. de verzameling van koppels die minstens k keer voorkomen in R. Een binaire multirelatie R wordt k-reflexief genoemd als de relatie R_k reflexief is. Een binaire multirelatie R is irreflexief als er geldt dat:

$$\forall o \in O : \omega_R(o, o) = 0. \tag{8.3}$$

Steunend op de definitie van multirelaties kunnen we de definitie van het relationeel model voor documenten geven.

Definitie 8.2 (Relationeel documentmodel)
We veronderstellen het bestaan van een conceptruimte $\mathcal{C}$. Een document (tekst) d is gedefinieerd als een verzameling van zinnen $S = \{s_1, ..., s_m\}$, waarbij een zin $s_i \subset \mathcal{M}(\mathcal{C} \times \mathcal{C})$ een binaire, irreflexieve, symmetrische en 1-transitieve multirelatie over $\mathcal{C}$ is. De verzameling van alle documenten wordt genoteerd als $\mathcal{D}$.

[2] Hoewel we dit niet expliciet vermelden, behandelen we steeds het geval van homogene relaties.

In wat volgt zullen we een multirelatie steeds als binair veronderstellen. Om die reden spreken we kortweg van een multirelatie, waar in principe een binaire multirelatie wordt bedoeld. In het relationeel model voor documenten wordt een document beschouwd als een verzameling van zinnen, waarbij elke zin een multirelatie van concepten is. Concepten zijn hierbij een collectie van zinvolle woorden die samen een semantisch geheel vormen. Het wordt opgemerkt dat volgens Definitie 8.2 geen uitspraak wordt gedaan over het soort relatie tussen concepten. Het enige wat we in het relationele model vastleggen, is het bestaan van een relatie tussen concepten die voorkomen binnen eenzelfde zin. Wanneer we een relationeel totaalbeeld willen van een document, kunnen we de relationele transformatie van dat document beschouwen.

Definitie 8.3 (Relationele transformatie)
Voor een document $d \in \mathcal{D}$ is de relationele transformatie gedefinieerd als een functie:

$$\psi : \mathcal{D} \rightarrow \mathcal{M}(\mathcal{C} \times \mathcal{C}) : d \mapsto \bigoplus_{i=1}^{|S|} s_i. \tag{8.4}$$

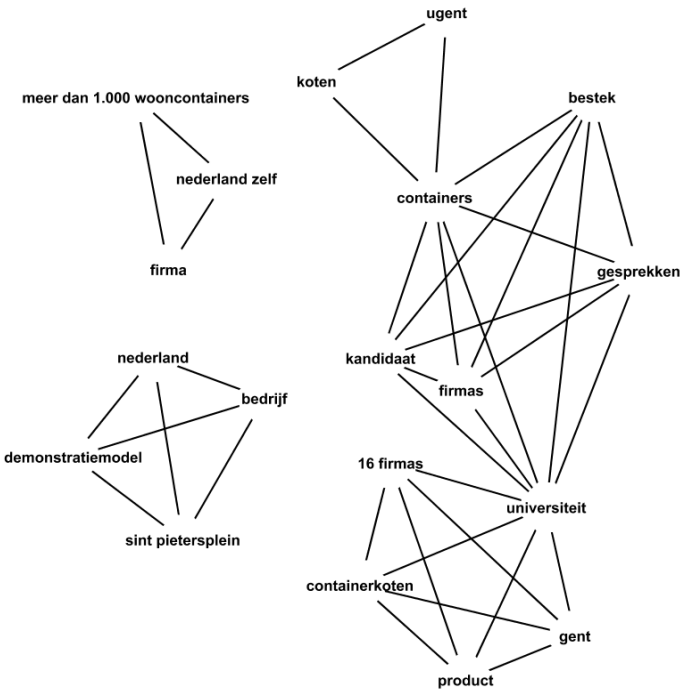
Dit betekent dat $\psi(d)$ gelijk is aan de som van alle zinnen.

Een belangrijke vraag die zich opdringt bij de introductie van een dergelijk nieuw model, is hoe een stuk tekst kan worden omgezet naar dit model. Een dergelijke omzetting is mogelijk door gebruik van de i.Know-*engine*. Deze *engine* is een gepatenteerde technologie van het bedrijf i.Know voor het omzetten van documenten naar een semi-gestructureerd model. Een bijzonder groot voordeel van deze technologie en tevens de grootste vernieuwing ten opzichte van bestaande methoden, is de context-onafhankelijkheid. In de literatuur zijn reeds vele intelligente methoden geïntroduceerd voor clustering van documenten op basis van semantische kennissystemen, maar het nadeel is hierbij steeds dat een dergelijke systeem context-afhankelijke concepten beschouwt. Daar waar klassieke NLP (Natural Language Processing) systemen een *top-down* aanpak gebruiken om termen te identificeren op basis van vooraf gedefinieerde thesauri, ontologieën of statistische modellen, gebruikt i.Know een *bottom-up* aanpak waarbij de SIE (Smart Indexing Engine) automatisch alle complexe termen in een tekst identificeert, los van hun lengte en hun semantische complexiteit. De i.Know-technologie maakt hierbij gebruik van de relationele structuur van een taal om alle concept-relatie-concept eenheden in een document automatisch te identificeren. Daar waar concepten bij een ontologie dus vooraf gekend dienen te zijn, biedt de i.Know-technologie een methode waarbij concepten dynamisch ontgonnen worden uit een tekst. Een concept is bijgevolg steeds een relevante combinatie van woorden, eerder dan een willekeurig q-gram. Deze eigenschap heeft als gevolg dat de frequentie van concepten doorheen een verzameling van documenten een indicatie vormt van de relevantie van dit concept. Dit verklaart meteen waarom we zinnen modelleren als multirelaties, eerder dan als gewone relaties: door de relationele transformatie (d.i. het sommeren van zinnen) kunnen relevante concepten, en bij uitbreiding relevante koppels van

concepten, snel worden gevonden door een snede (Hoofdstuk 1) van de resulterende multiverzameling of multirelatie te nemen. We benadrukken dat dit niet in de vorm van een stelling te formuleren valt. We aanvaarden dit eerder als een axiomatisch beginsel. De correctheid van dit beginsel moet experimenteel worden vastgesteld. Alvorens de eigenschappen van het nieuwe model verder te onderzoeken, beschouwen we eerst een voorbeeld.

Voorbeeld 8.1
Beschouw het volgende stuk tekst, bestaande uit vijf zinnen:

"UGent beoordeelt koten in containers. In Gent komen 16 firma's van containerkoten hun product voorstellen aan de universiteit. Een bedrijf uit Nederland heeft een demonstratiemodel meegebracht en op het Sint-Pietersplein gezet. In Nederland zelf bouwde de firma al meer dan 1.000 wooncontainers. Na de gesprekken zal de universiteit een bestek opmaken en kunnen de firma's zich kandidaat stellen om de containers te bouwen."



Figuur 8.2: Relationele transformatie van een document

*Het verwerken van de eerste zin met de i.Know technologie geeft het volgende resultaat "**ugent** beoordeelt **koten** in **containers**". Hierbij zijn concepten in vet gezet. Irrelevante woorden worden door de SIE weggelaten. In deze zin worden drie concepten en twee relaties gevonden. Op basis van deze input wordt een symmetrische en 1-transitieve multirelatie geconstrueerd door elk*

concept met elk ander concept in relatie te brengen. De multirelatie duidt aan dat twee concepten in dezelfde zin voorkomen. Wanneer we dit voor elke zin doen, kunnen we alle zinnen sommeren door gebruik te maken van de operator $\oplus$ voor multiverzamelingen, hetgeen ons de relationele transformatie van het document geeft. Het resultaat van deze transformatie wordt getoond in Figuur 8.2. Merk op dat uit Figuur 8.2 blijkt dat $\psi(d)$ niet noodzakelijk 1-transitief is (bvb.: concepten "ugent" en "bestek" staan niet met elkaar in relatie).

Laat ons nu een aantal definities en operatoren introduceren in verband met dit nieuwe model.

Definitie 8.4
Gegeven een document $d \in \mathcal{D}$ en een koppel van concepten $(c_1, c_2) \in \mathcal{C}^2$. Het koppel (c_1, c_2) behoort tot d, genoteerd als $(c_1, c_2) \in d$ als er geldt dat:

$$(c_1, c_2) \in \psi(d). \tag{8.5}$$

Definitie 8.5 (Concepten van een document)
Gegeven een document $d \in \mathcal{D}$ en zijn relationele transformatie $\psi(d)$. De concepten van d zijn gegeven als een multiverzameling $\mathcal{C}_d$ waarvoor er geldt dat:

$$\forall c_1 \in \mathcal{C} : \omega_{\mathcal{C}_d}(c_1) = \sum_{c_2 \in \mathcal{C}} \omega_{\psi(d)}(c_1, c_2). \tag{8.6}$$

Definitie 8.5 impliceert dat de concepten uit d worden voorgesteld als een multiverzameling, waarbij voor een concept c, de multipliciteit $\omega_{\mathcal{C}_d}(c)$ aangeeft hoeveel koppels er zijn in $\psi(d)$, die c als eerste concept hebben[3]. De voorstelling van de concepten van een document als multiverzameling bewaart dus de informatie over de frequentie van een concept binnen het document. In wat volgt zullen we willen verifiëren of de componenten van een koppel van concepten (mogelijks) voorkomen in de conceptmultiverzameling van een document. Daarom voeren we de volgende definities in.

Definitie 8.6
Voor een willekeurig document $d \in \mathcal{D}$ met conceptruimte $\mathcal{C}$ definiëren we:

$$\forall (c_1, c_2) \in \mathcal{C}^2 : (c_1, c_2) \mathrel{\widehat{\in}} d \Leftrightarrow (c_1 \in \mathcal{C}_d \wedge c_2 \in \mathcal{C}_d)\,. \tag{8.7}$$

Definitie 8.7
Voor een willekeurig document $d \in \mathcal{D}$ met conceptruimte $\mathcal{C}$ en een evaluator $E_{\mathcal{C}}$ definiëren we:

$$\forall (c_1, c_2) \in \mathcal{C}^2 : (c_1, c_2) \mathrel{\widetilde{\in}} d \Leftrightarrow \left(\exists (c_1', c_2') \mathrel{\widehat{\in}} d : \mathscr{B}(E_{\mathcal{C}}(c_1, c_1')) \wedge \mathscr{B}(E_{\mathcal{C}}(c_2, c_2'))\right). \tag{8.8}$$

waarbij $\mathscr{B}$ een maximaal beslissingsmodel is (d.i. drempelwaarde $z = 0$).

[3]We kijken enkel naar het eerste concept vermits de multirelaties onder beschouwing steeds symmetrisch zijn (Definitie 8.3).

Op basis van de $\in$-operator kunnen we relationele selectie van documenten definiëren.

Definitie 8.8 (Relationele selectie van documenten)
Gegeven een koppel van concepten $(c_1, c_2) \in \mathcal{C}$, dan is de relationele selectie van documenten onder (c_1, c_2) gedefinieerd als een verzameling documenten:

$$\mathcal{D}_{(c_1,c_2),\in} = \{d \in \mathcal{D} | (c_1, c_2) \in d\}. \tag{8.9}$$

Het koppel (c_1, c_2) wordt een zoekpatroon (of kortweg een patroon) genoemd.

Naast de relationele selectie kunnen we operatoren voor selectie definiëren met zwakkere selectiecriteria.

Definitie 8.9 (Conceptuele selectie van documenten)
Gegeven een koppel van concepten $(c_1, c_2) \in \mathcal{C}$, dan is de conceptuele selectie van documenten onder (c_1, c_2) gedefinieerd als een verzameling documenten:

$$\mathcal{D}_{(c_1,c_2),\,\widehat{\in}} = \left\{d \in \mathcal{D} | (c_1, c_2) \,\widehat{\in}\, d\right\}. \tag{8.10}$$

Het koppel (c_1, c_2) wordt een zoekpatroon (of kortweg een patroon) genoemd.

Definitie 8.10 (Possibilistische selectie van documenten)
Gegeven een koppel van concepten $(c_1, c_2) \in \mathcal{C}$, dan is de possibilistische selectie van documenten onder (c_1, c_2) gedefinieerd als een verzameling documenten:

$$\mathcal{D}_{(c_1,c_2),\,\widetilde{\in}} = \left\{d \in \mathcal{D} | (c_1, c_2) \,\widetilde{\in}\, d\right\} \tag{8.11}$$

Het koppel (c_1, c_2) wordt een zoekpatroon (of kortweg een patroon) genoemd.

Op dit punt kunnen enkele interessante verbanden worden gelegd met voorgaande hoofdstukken. We stellen vast dat elk concept c een string is. We kunnen voor elke conceptruimte stellen dat $\mathcal{C} \subset \mathcal{S}$. Dit betekent dat we zowel een syntactische evaluator voor strings (d.i. $E_{\mathcal{S}}$ of $E^*_{\mathcal{S}}$) als een semantische evaluator voor strings (d.i. $E_{\mathcal{S},R}$) kunnen gebruiken voor de possibilistische selectie van documenten (Hoofdstuk 6). We stellen ook vast dat de possibilistische selectie gebruik maakt van een evaluatieketting waarin twee tweewaardige evaluatoren worden gebruikt (Hoofdstuk 2). Op basis van de relationele selectie van documenten is het mogelijk afhankelijkheden tussen zoekpatronen te definiëren.

Definitie 8.11 (Afhankelijke patronen)
Twee patronen (c_1, c_2) en (c'_1, c'_2) worden afhankelijk genoemd, genoteerd als $(c_1, c_2) \sim (c'_1, c'_2)$, als er geldt dat:

$$\mathcal{D}_{(c_1,c_2),\in} \cap \mathcal{D}_{(c'_1,c'_2),\in} \neq \emptyset. \tag{8.12}$$

Afhankelijkheid van twee patronen duidt erop dat er minstens één document bestaat waarin beide patronen voorkomen. Het is mogelijk om zwakkere afhankelijkheden te definiëren op basis van $\widehat{\in}$ of $\widetilde{\in}$. Dit doen we hier niet aangezien dergelijke afhankelijkheden niet worden gebruikt in onze aanpak voor coreferentiebepaling van documenten. Voor een collectie van patronen is het mogelijk een afhankelijkheidsmatrix te definiëren.

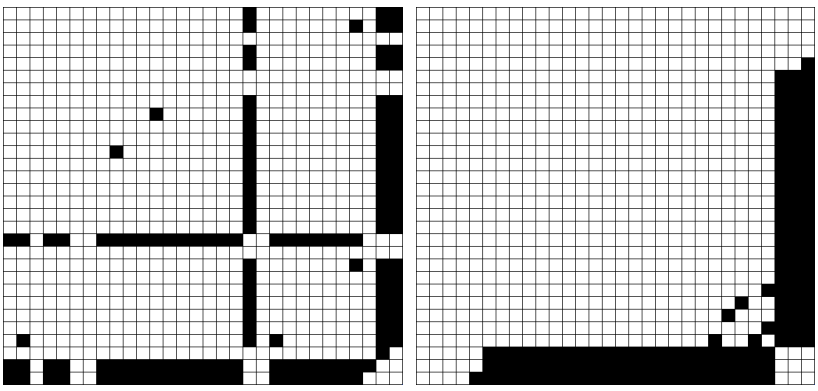
Definitie 8.12 (Afhankelijkheidsmatrix)
Gegeven een vector van patronen $\mathbf{v}$*, dan is de afhankelijkheidsmatrix voor* $\mathbf{v}$ *een* $|\mathbf{v}| \times |\mathbf{v}|$ *matrix* $\mathbf{M_v}$ *zodat er geldt:*

$$\mathbf{M_v}(i,j) = \begin{cases} 1 & \textbf{\textit{als}} \quad \mathbf{v}_i \sim \mathbf{v}_j \\ 0 & \textbf{\textit{anders}}. \end{cases} \tag{8.13}$$

Wanneer alle patronen in $\mathbf{v}$ onafhankelijk zijn, dan is $\mathbf{M_v}$ gelijk aan de eenheidsmatrix $\mathbb{I}_{|\mathbf{v}|}$. Beschouwen we een vector van patronen $\mathbf{v}$ en de bijhorende afhankelijkheidsmatrix $\mathbf{M_v}$, dan bestaat er steeds een permutatie van $\mathbf{v}$, genoteerd als $\mathbf{v}^*$, zodat er geldt:

$$\forall j \in \{1, ..., |\mathbf{v}| - 1\} : \sum_{i=1}^{|\mathbf{v}|} \mathbf{M_{v^*}}(i,j) \geq \sum_{i=1}^{|\mathbf{v}|} \mathbf{M_{v^*}}(i,j+1). \tag{8.14}$$

Dit betekent dat we een vector van patronen kunnen ordenen zodat de patronen met het meeste afhankelijkheden eerst staan. Een voorbeeld van een afhankelijkheidsmatrix $\mathbf{M_v}$ en de bijhorende $\mathbf{M_{v^*}}$ wordt getoond in Figuur 8.3, waar afhankelijkheid wordt voorgesteld door een wit vakje en onafhankelijkheid door een zwart vakje. De permutatie van $\mathbf{v}$ tot $\mathbf{v}^*$ zorgt ervoor dat het aantal onafhankelijkheden op de kolommen van $\mathbf{M_{v^*}}$ toenemen van links naar rechts.



Figuur 8.3: Afhankelijkheidsmatrix $\mathbf{M_v}$ (links) en de afgeleide $\mathbf{M_{v^*}}$ (rechts)

Wanneer informatie nodig is over de mate waarin twee patronen afhankelijk zijn, volstaat de definitie van afhankelijkheid niet en is er nood aan een graad van afhankelijkheid.

Definitie 8.13 (Afhankelijkheidsgraad)
Gegeven twee patronen (c_1, c_2) en (c'_1, c'_2). De graad van afhankelijkheid tussen deze patronen is gedefinieerd als:

$$\text{dep}((c_1, c_2), (c'_1, c'_2)) = \frac{\left|\mathcal{D}_{(c_1,c_2),\in} \cap \mathcal{D}_{(c'_1,c'_2),\in}\right|}{\min\left(|\mathcal{D}_{(c_1,c_2),\in}|, |\mathcal{D}_{(c'_1,c'_2),\in}|\right)}. \tag{8.15}$$

De graad van afhankelijkheid is gelegen in het eenheidsinterval $[0, 1]$ en er geldt dat:

$$\text{dep}((c_1, c_2), (c'_1, c'_2)) = \text{dep}((c'_1, c'_2), (c_1, c_2)). \tag{8.16}$$

Er geldt ook dat:

$$(c_1, c_2) \sim (c'_1, c'_2) \Rightarrow \text{dep}((c_1, c_2), (c'_1, c'_2)) > 0. \tag{8.17}$$

Met het nieuwe tekstmodel voorhanden zullen we in de volgende secties bestuderen hoe dit nieuwe model kan worden gebruikt voor de innovatie van tekstclustering.

8.4 Evaluatie van documenten

In deze sectie introduceren we evaluatoren voor documenten. In de context van documenten worden de beschreven entiteiten ook onderwerpen genoemd. In wat volgt zullen we de termen 'entiteit' en 'onderwerp' daarom als equivalent beschouwen. De objecten zijn in dit hoofdstuk documenten. In Hoofdstuk 2 is aangehaald dat objecten niet altijd rechtstreeks bruikbaar zijn voor vergelijking. In de context van beelden is er bijvoorbeeld meestal een transformatie nodig naar een objectruimte waarin de getransformeerde objecten kunnen worden vergeleken. Dergelijke transformaties kunnen worden beschouwd als meetprocessen $\mathscr{M}$, waardoor transformaties onderhevig kunnen zijn aan meetimperfecties. In het geval van documenten hebben we reeds een transformatiefunctie gedefinieerd (d.i. de relationele transformatiefunctie ψ). Het merendeel van de patronen in $\psi(d)$ biedt echter voornamelijk detailinformatie over de entiteit die door d wordt beschreven, net zoals het merendeel van de informatie in een beeld detailinformatie is. Om die reden willen we voor documenten een bijkomende transformatie doorvoeren die een selectie maakt van de meest relevante patronen.

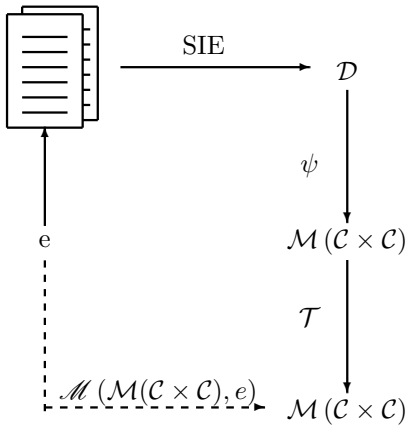
Definitie 8.14
Gegeven een documentruimte $\mathcal{D}$, een conceptruimte $\mathcal{C}$ en een relationele transformatiefunctie ψ, dan definiëren we een functie:

$$\mathcal{T} : \mathcal{M}(\mathcal{C} \times \mathcal{C}) \to \mathcal{M}(\mathcal{C} \times \mathcal{C}) \tag{8.18}$$

waarvoor er geldt dat:

$$\forall d \in \mathcal{D} : \mathcal{T}(\psi(d)) \subseteq \psi(d). \tag{8.19}$$

Figuur 8.4 toont het resulterende meetproces. Vertrekkende van een entiteit e resulteert de vraag "Hoe zou u e beschrijven?" in een tekstuele beschrijving van e die door de SIE kan worden omgezet naar een object in het universum $\mathcal{D}$ (d.i. een document). Elk document uit $\mathcal{D}$ kan worden getransformeerd door de relationele transformatiefunctie ψ te gebruiken. Het resultaat is een multirelatie over de conceptruimte $\mathcal{C}$. De transformatiefunctie $\mathcal{T}$ reduceert deze multirelatie door niet relevante patronen te verwijderen.



Figuur 8.4: Meting van entiteiten ontbonden in transformaties

Terugkerend naar het voorbeeld gegeven door Figuur 2.3, stellen we vast dat de transformaties die worden getoond in Figuur 2.3 contextonafhankelijk zijn. Dit betekent dat de transformatie kan gebeuren door enkel het te transformeren object (in Figuur 2.3 een foto) te beschouwen. Deze veronderstelling gaat niet op voor het geval van documenten. Ons onderzoek heeft geleerd dat de selectie van relevante patronen door $\mathcal{T}$ voor een document d gebeurt door observatie van andere documenten in een gegeven collectie van documenten D. Anders gezegd: twee documenten d_1 en d_2 worden steeds vergeleken in een context $D \subset \mathcal{D}$. Dit vormt een belangrijke uitdaging voor het vergelijken van documenten.

Als we veronderstellen dat dit probleem kan worden opgelost en dat $\mathcal{T}$ een gekende functie is, kunnen we een evaluator voor documenten definiëren als volgt.

Definitie 8.15 (Evaluator voor documenten)
Een evaluator voor documenten is gedefinieerd als:

$$E_{\mathcal{D}} : \mathcal{D}^2 \to \mathcal{F}(\mathbb{B}) : (d_1, d_2) \mapsto E_{\mathcal{M}(\mathcal{C}\times\mathcal{C})}\Big(\mathcal{T}(\psi(d_1)), \mathcal{T}(\psi(d_2))\Big). \tag{8.20}$$

De evaluator $E_{\mathcal{M}(C\times C)}$ is een hierbij een evaluator voor collecties (Hoofdstuk 4). We herhalen hier dat een dergelijke evaluator in twee stappen werkt. Eerst wordt een leximax-optimale één-op-één afbeelding ι geconstrueerd op

basis van een evaluator $E_{\mathcal{C}\times\mathcal{C}}$. Deze afbeelding leidt tot een multiverzameling van possibilistische waarheidswaarden en deze multiverzameling wordt verwerkt door de combinatiefunctie $S_{\gamma^{T,F}}$, waarbij γ^T en γ^F aan welbepaalde voorwaarden moeten voldoen (Hoofdstuk 4). Laat ons nu, gelet op het feit dat $\mathcal{T}$ onbekend is, enkele veronderstellingen maken die $E_{\mathcal{D}}$ aanzienlijk vereenvoudigen. Veronderstel dat $E_{\mathcal{C}\times\mathcal{C}}$ een tweewaardige evaluator is (Hoofdstuk 2). In dit geval geldt er dat:

$$\Big(E_{\mathcal{C}\times\mathcal{C}}\Big(\,(c_1,c_2)\,,(c_1',c_2')\,\Big)\neq(0,1)\Big)\Leftrightarrow\Big((c_1,c_2)=(c_1',c_2')\Big). \tag{8.21}$$

Veronderstel[4] ook dat $S_{\gamma^{T,F}}=\widetilde{\vee}$. In dit geval geldt er voor twee willekeurige documenten $d_1\in\mathcal{D}$ en $d_2\in\mathcal{D}$ dat:

$$E_{\mathcal{D}}(d_1,d_2)=\begin{cases}(1,0) & \textbf{als} \quad \mathcal{T}(\psi(d_1))\cap\mathcal{T}(\psi(d_2))\neq\emptyset\\ (0,1) & \textbf{anders}.\end{cases} \tag{8.22}$$

Als we veronderstellen dat er een relatie $R^{(\mathcal{D})}\subset\mathcal{C}\times\mathcal{C}$ bestaat zodat er geldt dat:

$$\forall d\in\mathcal{D}:\mathcal{T}(\psi(d))=\psi(d)\cap R^{(\mathcal{D})} \tag{8.23}$$

dan kunnen we de voorwaarde uit (8.22) waaronder $E_{\mathcal{D}}(d_1,d_2)=(1,0)$, herschrijven als:

$$\Big(\psi(d_1)\cap R^{(\mathcal{D})}\Big)\cap\Big(\psi(d_2)\cap R^{(\mathcal{D})}\Big)\neq\emptyset \tag{8.24}$$

hetgeen ook kan worden geschreven als:

$$\exists(c_1,c_2)\in R^{(\mathcal{D})}:d_1\in\mathcal{D}_{(c_1,c_2),\in}\wedge d_2\in\mathcal{D}_{(c_1,c_2),\in}. \tag{8.25}$$

Bijgevolg geldt er dat:

$$E_{\mathcal{D}}(d_1,d_2)=\begin{cases}(1,0) & \textbf{als} \quad \exists(c_1,c_2)\in R^{(\mathcal{D})}:(d_1,d_2)\in\left(\mathcal{D}_{(c_1,c_2),\in}\right)^2\\ (0,1) & \textbf{anders}.\end{cases} \tag{8.26}$$

Deze formulering van $E_{\mathcal{D}}$ is belangrijk om twee redenen. Ten eerste volgt er voor een patroon $(c_1,c_2)\in R^{(\mathcal{D})}$ dat $\mathcal{D}_{(c_1,c_2),\in}$ een verzameling van coreferente documenten is. Als $R^{(\mathcal{D})}$ gekend zou zijn, kunnen onmiddellijk collecties van coreferente documenten worden afgeleid, zonder $E_{\mathcal{D}}$ nog expliciet in rekening te brengen. Interessant genoeg zien we nu een verband met $\cap$-coreferentie (Hoofdstuk 2). Meer bepaald is het zo dat, onder deze veronderstellingen, een document d tot meerdere clusters kan behoren. In de context van documenten is dit niet onlogisch aangezien een document inderdaad meerdere entiteiten (d.z. onderwerpen) kan beschrijven. Echter, in Hoofdstuk 5 is aangehaald dat in sommige situaties een $\cap$-coreferentieprobleem goed kan worden benaderd als een gewoon coreferentieprobleem. We zullen daarom veronderstellen dat een document steeds exact één onderwerp beschrijft, waardoor $\cap$-coreferentie wordt

[4]Deze veronderstelling is toegelaten onder de voorwaarden voor γ^T en γ^F afgeleid in Hoofdstuk 4.

herleid tot gewone coreferentie. In het geval waarbij een document meerdere entiteiten beschrijft, zullen we onze benadering handhaven door een hoofdentiteit aan te duiden. De benadering tot gewone coreferentie betekent dat we zullen proberen om een collectie van patronen $R^{(\mathcal{D})}$ te bepalen en op basis daarvan een partitie van $\mathcal{D}$ af te leiden. De partitieklassen komen dan overeen met clusters. Hiervoor zullen we in Sectie 8.5 een methode construeren die het aantal partitieklassen (clusters) schat. In Sectie 8.6 beschrijven we hoe $R^{(\mathcal{D})}$ kan worden bepaald en hoe de kwaliteit van de overeenkomstige partitieklassen (clusters) kan worden geschat, om zo te komen tot een optimale partitie.

Ten tweede is het mogelijk om de selectie van documenten $\mathcal{D}_{(c_1,c_2),\in}$ in (8.26) te vervangen door $\mathcal{D}_{(c_1,c_2),\,\widehat{\in}}$ of $\mathcal{D}_{(c_1,c_2),\,\widetilde{\in}}$, hetgeen alternatieve evaluatoren bepaalt.

Definitie 8.16 (op**-evaluator voor documenten)**
Een op*-evaluator voor documenten met* op $\in \{\in, \widehat{\in}, \widetilde{\in}\}$ *is gedefinieerd als:*

$$E^{\text{op}}_{\mathcal{D}} : \mathcal{D}^2 \to \mathcal{F}(\mathbb{B}) : (d_1, d_2) \mapsto E^{\text{op}}_{\mathcal{M}(\mathcal{C}\times\mathcal{C})}(\mathcal{T}(\psi(d_1)), \mathcal{T}(\psi(d_2))) \tag{8.27}$$

zodat:

$$E^{\text{op}}_{\mathcal{D}}(d_1, d_2) = \begin{cases} (1,0) & \textbf{\textit{als}} \quad \exists (c_1, c_2) \in R^{(\mathcal{D})} : (d_1, d_2) \in \left(\mathcal{D}_{(c_1,c_2),\text{op}}\right)^2 \\ (0,1) & \textbf{\textit{anders}}. \end{cases} \tag{8.28}$$

Stelling 8.1
Gegeven een documentruimte $\mathcal{D}$, dan geldt er dat:

$$\forall (d_1, d_2) \in \mathcal{D} : E^{\in}_{\mathcal{D}}(d_1, d_2) \leq E^{\widehat{\in}}_{\mathcal{D}}(d_1, d_2) \leq E^{\widetilde{\in}}_{\mathcal{D}}(d_1, d_2) \tag{8.29}$$

Bewijs. Het bewijs volgt uit het feit dat er voor een patroon $(c_1, c_2) \in \mathcal{C}^2$ geldt dat:

$$\mathcal{D}_{(c_1,c_2),\in} \subseteq \mathcal{D}_{(c_1,c_2),\,\widehat{\in}} \subseteq \mathcal{D}_{(c_1,c_2),\,\widetilde{\in}}\,. \tag{8.30}$$

□

Op basis van deze alternatieve evaluatoren kunnen we een verzameling van coreferente documenten gaan bepalen als:

$$\mathcal{D}_{(c_1,c_2),\text{op}}. \tag{8.31}$$

In Sectie 8.7 wordt onderzocht wat de impact van het gebruik van deze alternatieve evaluatoren is.

Tot slot van deze sectie bespreken we de compatibiliteit van deze aanpak met het raamwerk voor complexe objecten. Beschouw een complex universum $O = U_1 \times ... \times U_n$ zodat minstens één van de atomaire universa gelijk is aan $\mathcal{D}$. Kunnen we de zonet beschreven aanpak voor documenten dan toepassen in het possibilistisch raamwerk voor complexe objecten uit Hoofdstuk 7? Het antwoord op deze vraag is ja. In onze possibilistische aanpak voor complexe objecten schrijven we voor dat elk atomair universum U_i moet worden voorzien

van een evaluator E_{U_i}. De bepaling van deze evaluator gebeurt hierbij volledig op de i^{de} projectie van het complex universum. De evaluator $E_{\mathcal{D}}$ kan worden afgeleid en kan als een deelevaluator worden gebruikt in het raamwerk voor complexe objecten.

8.5 Schatting van het aantal entiteiten

In deze sectie zal een nieuwe methode worden geïntroduceerd om, gegeven een verzameling van documenten, een schatting te maken van het aantal clusters dat in deze verzameling aanwezig is. In termen van coreferente documenten wil dit zeggen: als elk document een entiteit beschrijft, hoeveel verschillende entiteiten worden dan door alle documenten samen beschreven? Volgende notaties worden in deze sectie gebruikt. We veronderstellen een verzameling van documenten $D = \{d_1, ..., d_n\}$ en een referentiefunctie ρ (Definitie 2.2). We herhalen dat we de termen 'entiteit' en 'onderwerp' als equivalent beschouwen. Het universum van entiteiten blijven we noteren als $\mathcal{E}$. Gegeven een verzameling van documenten D, beschouw de multiverzameling $\mathcal{E}_D \subset \mathcal{M}(\mathcal{E})$ zodat:

$$\forall e \in \mathcal{E} : \omega_{\mathcal{E}_D}(e) = \Big|\{d | d \in D \wedge \rho(d) = e\}\Big|. \tag{8.32}$$

Het natuurlijke getal $\omega_{\mathcal{E}_D}(e)$ wordt de kardinaliteit van e in D genoemd. Het aantal clusters in een verzameling van documenten D is bijgevolg gelijk aan $|\,(\mathcal{E}_D)_1\,|$, d.i. de kardinaliteit van de 1-snede van $\mathcal{E}_D$.

In de literatuur over clustering vinden we twee belangrijke families van methoden voor het bepalen van het aantal clusters. Een eerste familie van methoden berekent het aantal clusters als het beeld van het aantal documenten onder een (heuristische) functie. Een voorbeeld hiervan vinden we in [121]:

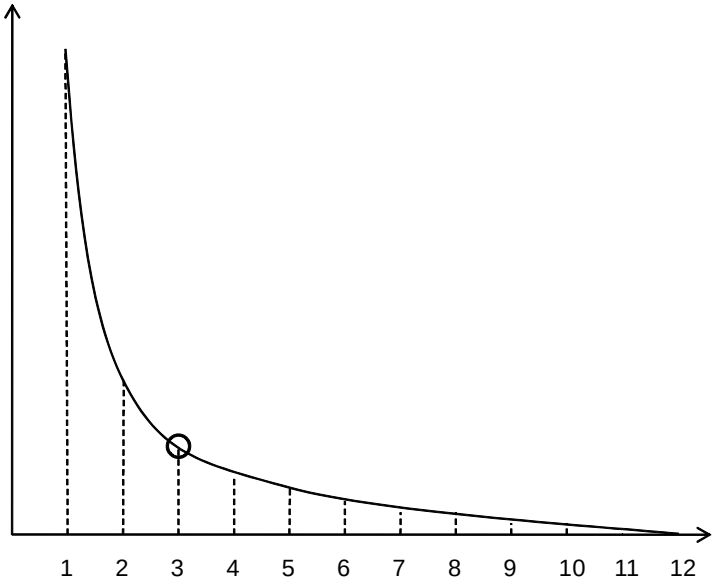
$$\sqrt{\left(\frac{|D|}{2}\right)}. \tag{8.33}$$

Dergelijke methoden hebben een constante complexiteit, maar hun accuraatheid kan ernstig in twijfel worden getrokken. Een nadeel van deze methode is verder dat het aantal clusters een constante is in functie van het aantal documenten. Bij een tweede familie van methoden wordt er verondersteld dat het te gebruiken clusteralgoritme gekend is en dat het aantal clusters een parameter van dit algoritme is [122, 123, 124, 125, 126]. Deze methoden stellen dat het aantal clusters wordt begrensd door het aantal documenten, d.i.:

$$|\,(\mathcal{E}_D)_1\,| \in \{1, ..., |D|\}. \tag{8.34}$$

De clusters kunnen dan worden bepaald voor verschillende waarden voor het aantal clusters. Bij elke berekening wordt een vooraf bepaald foutcriterium gemeten zoals de gemiddelde intra-clusterfout. Op basis van deze metingen kiest men dan een optimaal aantal clusters. De gemiddelde intra-clusterfout heeft bijvoorbeeld typisch een $1/n$ verloop in functie van het aantal clusters n

(op een schaalfactor na). Het keuzecriterium bestaat dan uit het zoeken naar een optimale richtingscoëfficiënt voor de raaklijn, bijvoorbeeld door het grootste aantal clusters te nemen zodat de absolute waarde van de richtingscoëfficiënt onder een drempelwaarde ligt. Een voorbeeld wordt getoond in Figuur 8.5, waarbij het cirkeltje het optimaal aantal clusters aangeeft.



Figuur 8.5: Bepaling van $|\,(\mathcal{E}_D)_1\,|$ op basis van de intra-clusterfout

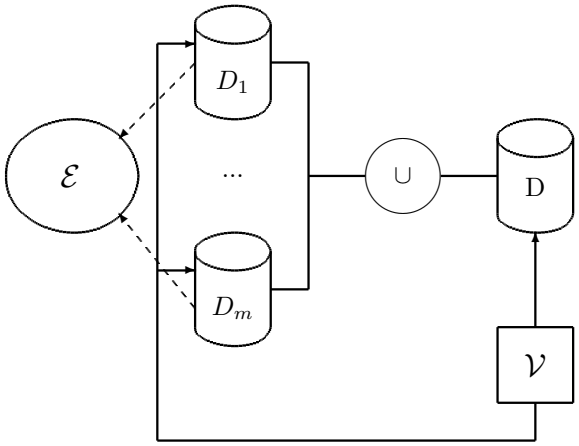
Deze familie van technieken is uitgebreid onderzocht in de literatuur en het is aangetoond dat de resultaten behoorlijk zijn in termen van accuraatheid. Het geschat aantal clusters is zeker niet altijd constant in functie van het aantal documenten. Echter, enkele nadelen worden ook bij deze familie vastgesteld. Ten eerste hebben de meeste clusteralgoritmen een behoorlijke complexiteit. In de context van documentclustering moeten klassieke clusteralgoritmen ook altijd worden gecombineerd met dimensiereductie [127]. Men kan zich dus de vraag stellen of het herhaaldelijk uitvoeren van deze algoritmen[5] haalbaar is in praktische situaties. Ten tweede hangt de kwaliteit van de schatting van het aantal clusters volledig af van de kwaliteit van het gebruikte algoritme. Een slechte keuze van het clusteralgoritme impliceert dus een slechte schatting van het aantal clusters. Een opvallende vaststelling die geldt voor zowel het schatten van het aantal clusters als voor het clusteren zelf, is dat veel methoden gebruik maken van klassieke technieken uit de *datamining*. Dit ondanks het feit dat documenten een aantal uitgesproken eigenschappen hebben in vergelijking met de data waarvoor deze klassieke technieken meestal worden gebruikt (hoge

[5]Dimensiereductie moet in principe slechts één keer worden uitgevoerd.

dimensie, semantiek ...). Daarom lijkt het ons nuttig te onderzoeken hoe het aantal clusters kan worden bepaald, los van het te gebruiken clusteralgoritme.

8.5.1 Competitieve entiteitsbeschrijving

De methode die hier wordt voorgesteld, is gebaseerd op de veronderstelling dat het genereren van documenten niet willekeurig verloopt, maar eerder gebeurt in een context van zogenaamde 'competitieve beschrijving' van entiteiten. De schematische voorstelling van deze competitieve context wordt nader toegelicht aan de hand van het schema in Figuur 8.6.



Figuur 8.6: Competitieve beschrijving van entiteiten

In dit schema wordt uitgegaan van m generatoren van data, die elk één databank beheren. De databank bevat objecten die in deze context een verzameling documenten D_i zijn. De verzameling van documenten D is dan bepaald als de unie van de verzamelingen D_i:

$$D = \bigcup_{i=1}^{m} D_i. \tag{8.35}$$

De verzameling D wordt geobserveerd door een verwerkingseenheid $\mathcal{V}$ die documenten in D kan raadplegen. Wanneer $\mathcal{V}$ een document $d \in (D \cap D_i)$ raadpleegt, waarvoor $\rho(d) = e$, dan stelt $\mathcal{B}(e)$ het bedrag voor dat $\mathcal{V}$ maximaal aan de beheerder van databank i wil geven, in ruil voor het raadplegen van d. We veronderstellen hierbij expliciet dat het te spenderen bedrag binnen een eindig tijdsinterval, eindig is. Zoniet kan $\mathcal{V}$ alle documenten raadplegen en valt het competitief karakter weg. Wanneer elke beheerder van data streeft naar maximalisatie van zijn/haar inkomen, dan kunnen we redeneren dat de generatie van D_i (en dus ook van D) aan bepaalde regels voldoet. De beheerder van databank D_i zal een schatting maken van de onderwerpen waar $\mathcal{V}$ interesse

voor heeft. Aangezien deze interesse recht evenredig is met $\mathcal{B}(e)$, wil dit zeggen dat:

$$\Pr(e|i) = \frac{\hat{\mathcal{B}}_i(e)}{\sum_{e' \in \mathcal{E}} \hat{\mathcal{B}}_i(e')} \tag{8.36}$$

de waarschijnlijkheid is dat databank i een document bevat over onderwerp e. Hierbij is $\hat{\mathcal{B}}_i(e)$ de schatting van $\mathcal{B}(e)$ door de beheerder van databank i. Anders gezegd, de waarschijnlijkheid dat databank i een document d over onderwerp e bevat, is gelijk aan de genormaliseerde interesse van de verwerkingseenheid $\mathcal{V}$ voor onderwerp e. Het bedrag $\mathcal{B}(e)$ kan verschillende fysieke vormen aannemen. In het kader van een website kan dit bijvoorbeeld het bezoeken van de site op zich zijn, waarbij elke website zijn/haar aantal bezoekers tracht te maximaliseren. Het competitief karakter komt dan voort uit het feit dat gebruikers van het WWW nooit alle websites zullen bezoeken. We willen hier benadrukken dat een dergelijk competitief model niet noodzakelijk is bij de opbouw van een verzameling van documenten. Eerder wordt gesteld dat een competitief karakter in vele praktische situaties geldt (website, onderzoeksdatabanken ...). Als de veronderstelling van een competitief karakter geldt, dan kan de hier beschreven aanpak worden gebruikt.

Enerzijds kunnen we nu veronderstellen dat $\mathcal{B}(e)$ een machtswet volgt. Als we de onderwerpen uit $\mathcal{E}$ rangschikken volgens $\mathcal{B}(e)$, dan zal $\mathcal{B}(e)$ exponentieel dalen. Anderzijds zal de concurrentie tussen beheerders van databanken tot gevolg hebben dat dit effect wordt versterkt. De spelende concurrentie zet beheerders van databanken aan tot een accurate schatting van $\mathcal{B}(e)$. Door de optimalisatie van hun inkomsten zullen verschillende databanken documenten over dezelfde onderwerpen bevatten. We kunnen dit alles modelleren door te stellen dat de interesse van $\mathcal{V}$ gehoorzaamt aan de Wet van Zipf [128]. Deze machtswet kan als volgt worden samengevat. Stel dat men aan een verzameling van onderwerpen een rangnummer i toekent die de onderwerpen ordent volgens dalende interesse (d.i., $\mathcal{B}(e_i) \geq \mathcal{B}(e_{i+1})$). In dat geval is de waarschijnlijkheid dat een willekeurig document over onderwerp e_i handelt gegeven door de Zipf-verdeling:

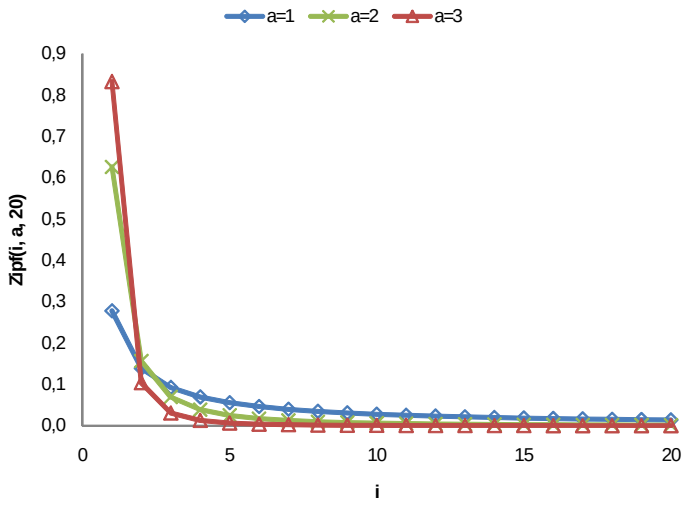
$$\forall i \in \{1, ..., b\} : \Pr(e_i) = \text{Zipf}(i, a, b) = \frac{1}{i^a H_{a,b}}. \tag{8.37}$$

Hierbij speelt $a \in \mathbb{R}_0^+$ de rol van een vormparameter en stelt $b \in \mathbb{N}_0$ het aantal entiteiten voor. $H_{a,b}$ is gelijk aan het b^{de} veralgemeend harmonisch getal:

$$H_{a,b} = \sum_{j=1}^{b} \frac{1}{j^a}. \tag{8.38}$$

Een belangrijke en gekende eigenschap van veralgemeende harmonische getallen is dat, voor $a > 1$, $H_{a,b}$ convergeert. Meer bepaald kan worden aangetoond dat er geldt:

$$\lim_{b \to \infty} H_{a,b} = \zeta(a) \tag{8.39}$$



Figuur 8.7: Zipf-verdeling

waarbij $\zeta(.)$ de Riemann-Zeta functie is. Een gevolg hiervan is dat de Zipf-verdeling ook convergeert voor $a > 1$ en $b \to \infty$. Figuur 8.7 toont de Zipf-verdeling ($b = 20$) voor drie verschillende waarden van a. Deze figuur illustreert duidelijk dat voor hogere waarden van a, de waarschijnlijkheidsmassa opschuift naar links. Met een dergelijk waarschijnlijkheidsmodel voor de interesse van $\mathcal{V}$ en bij uitbreiding de verdeling van documenten over onderwerpen, kunnen we de multiverzameling $\mathcal{E}_D$ bij benadering herconstrueren als volgt:

$$\omega_{\mathcal{E}_D}(e_i) = \lceil \text{Zipf}(i, a, b).|D| - 0.5 \rceil \tag{8.40}$$

waarbij de functie $\lceil x - 0.5 \rceil$ het reële getal x afrondt naar het dichtstbijzijnde natuurlijke getal. Dit is uiteraard een benadering aangezien kleine afrondingsfouten niet worden vermeden.

Wanneer we erin slagen om de parameters a en b van de Zipf-verdeling te bepalen, zodat de overeenkomstige waarschijnlijkheidsverdeling de verdeling van documenten in D over onderwerpen optimaal beschrijft, dan verkrijgen we automatisch een bepaling van het aantal onderwerpen en dus ook van het aantal clusters. We zullen een methode voorstellen die op basis van een benadering van de singletonclusters (Sectie 8.5.2) een optimaal aantal clusters bepaalt. Hiervoor analyseren we eerst de vormparameter a. Er is al vermeld dat wanneer $a > 1$, $\text{Zipf}(i, a, b)$ convergeert als $b \to \infty$. Een interessant bijkomend verschijnsel is dat de snelheid van de convergentie een exponentieel stijgend karakter heeft. Neem bijvoorbeeld $a = 2$ dan geldt er:

$$\text{Zipf}(1, 2, 5) - \text{Zipf}(1, 2, 10000) \approx 0.075 \tag{8.41}$$

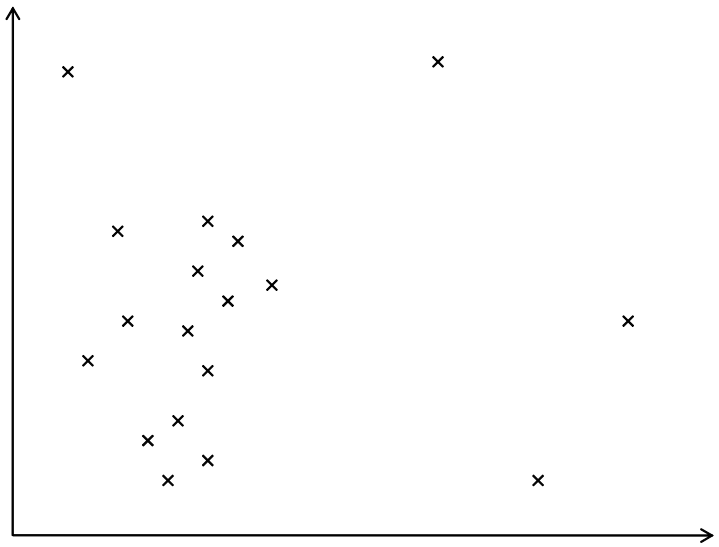
$$\text{Zipf}(1, 2, 10) - \text{Zipf}(1, 2, 10000) \approx 0.037. \tag{8.42}$$

Deze cijfers stellen dat bij $a = 2$, de waarschijnlijkheid dat een willekeurig document handelt over onderwerp e_1 bijna gelijk blijft, ongeacht het aantal documenten dat in totaal wordt beschouwd. Een dergelijk model is hoogst onrealistisch aangezien er steeds een relatief groot aantal onderwerpen bestaan waarvoor (weinig) documenten beschikbaar zijn. Met andere woorden, $a = 2$ levert geen realistisch model omdat er enorm veel waarschijnlijkheidsmassa verdeeld is over een zeer beperkt aantal onderwerpen. De Wet van Zipf kan dus enkel realistisch zijn als de Zipf-verdeling een zogenaamde *lange staart* heeft. Deze vaststelling wordt ook gemaakt in andere toepassingen van de Wet van Zipf, zoals de gebruiksfrequentie van woorden in een natuurlijke taal [128] en de economische studie van niche toepassingen. Ondanks het domein van a bestaan er dus slechts een zeer beperkt aantal waarden voor a die kunnen resulteren in een realistisch model. We zullen in een volgende sectie bespreken hoe we uit deze beperkte verzameling van waarden een goede waarde voor a kunnen kiezen.

8.5.2 Singletonclusters

In deze sectie bespreken we hoe, gegeven een verzameling van documenten D, parameters a en b bepaald kunnen worden, zodat de overeenkomstige Zipf-verdeling zo goed mogelijk de verdeling van documenten over entiteiten beschrijft. Laat ons eerst veronderstellen dat de waarde van parameter a gekend is. Dit betekent dat voor verschillende waarden van b, een reconstructie van $\mathcal{E}_D$ kan worden gemaakt door gebruik van (8.40). Om te weten welke van deze reconstructies de beste is, maken we gebruik van het aantal singletonclusters. Dit zijn clusters waarin slechts één document zit. Dergelijke clusters hebben de eigenschap dat ze in vrijwel elke situatie makkelijk herkenbaar zijn. Beschouw bijvoorbeeld de datapunten afgebeeld in Figuur 8.8, dan kunnen de vier singletonclusters vrijwel onmiddellijk worden herkend als de geïsoleerde datapunten. In de verschillende reconstructies van $\mathcal{E}_D$ kunnen we nu de singletonclusters gaan tellen en nagaan welke reconstructie het echte aantal singletonclusters het beste benadert. De waarde van b die overeenkomt met deze reconstructie is dan de schatting voor het aantal aanwezige clusters. Uiteraard kennen we het echte aantal singletonclusters niet, maar dit aantal is relatief eenvoudig te benaderen als volgt. Voor een gegeven document d, is $\mathcal{C}_d$ een multiverzameling van concepten. Door de eigenschap van concepten dat hun frequentie een maatstaf is voor relevantie, kunnen we stellen dat een document in een singletoncluster zit, als het geen enkel van zijn relevante concepten gemeen heeft met andere documenten. Veronderstel daarom dat voor elk document een drempelwaarde $w(d) \in \mathbb{N}$ gegeven is, zodat $(\mathcal{C}_d)_{w(d)}$ de verzameling van relevante concepten is. Dit wil zeggen, de relevante concepten van een document d, zijn deze die minstens $w(d)$ keer voorkomen in het document. Voor onze benadering veronderstellen we dat d behoort tot een singletoncluster als er geldt dat:

$$\forall d' \in D \backslash \{d\} : (\mathcal{C}_{d'})_{w(d')} \cap (\mathcal{C}_d)_{w(d)} = \emptyset. \tag{8.43}$$



Figuur 8.8: Datacollectie met singletonclusters

Algoritme 8.1 vat de voorgestelde methode samen in pseudocode.

Algoritme 8.1 estimate(D)

$\hat{\sigma} \leftarrow \left|\left\{d | d \in D \wedge \forall d' \in D \backslash \{d\} : (\mathcal{C}_{d'})_{w(d')} \cap (\mathcal{C}_d)_{w(d)} = \emptyset\right\}\right|$
for $i \in \{1, ..., |D|\}$ **do**
  if $\forall j \in \{1, ..., i\} : \lceil \text{Zipf}(j, a, i)|D| - 0.5 \rceil > 0$ **then**
    $\sigma_i \leftarrow |\{j | j \in \{1, ..., i\} \wedge \lceil \text{Zipf}(j, a, i)|D| - 0.5 \rceil = 1\}|$
  end if
end for
$|\widehat{(\mathcal{E}_D)_1}| \leftarrow \arg\min_i |\hat{\sigma} - \sigma_i|$
return $|\widehat{(\mathcal{E}_D)_1}|$

Voor het berekenen van de drempelwaarden redeneren we als volgt. Uit metingen blijkt dat 96% van de concepten uit een conceptverzameling $\mathcal{C}_d$ een multipliciteit strikt kleiner dan 3 hebben. Deze metingen bevestigen de aanname dat de multipliciteit van een concept een indicatie is van relevantie. Ook blijkt dat in 75% van de gevallen de maximale multipliciteit in een conceptmultiverzameling $\mathcal{C}_d$ groter dan of gelijk is aan 3. Uit deze metingen leiden we af dat de relevante concepten deze zijn die een maximale multipliciteit hebben, behalve wanneer de maximale multipliciteit te hoog is. In dat geval kunnen relevante concepten een lagere multipliciteit hebben. De metingen leren ons dat een grens voor hoge multipliciteit op 3 kan worden gelegd. Wanneer voor

een document d de maximale multipliciteit:

$$m = \max_{c \in \mathcal{C}} \omega_{\mathcal{C}_d}(c) \tag{8.44}$$

groter is dan of gelijk aan 3, zullen we als drempelwaarde $m-1$ kiezen. Zoniet kiezen we m. De drempelwaarden $w(d)$ worden dan als volgt berekend:

$$w(d) = \begin{cases} \max_{c \in \mathcal{C}} \omega_{\mathcal{C}_d}(c) - 1 & , \max_{c \in \mathcal{C}} \omega_{\mathcal{C}_d}(c) \geq 3 \\ \max_{c \in \mathcal{C}} \omega_{\mathcal{C}_d}(c) & , \max_{c \in \mathcal{C}} \omega_{\mathcal{C}_d}(c) < 3. \end{cases} \tag{8.45}$$

Voor de bepaling van de vormparameter a, blijkt uit experimenten (Sectie 8.7) dat een goede waarde voor a de grootste waarde is waarvoor het gemiddelde van $|(\mathcal{E}_D)_1|$ onder constante $|D|$ significant groter is dan 0. Significantie kan hierbij worden bepaald aan de hand van een t-toets. Dit kan worden verklaard door op te merken dat een overschatting van a tot gevolg heeft dat de bovengrens voor de schatting van het aantal clusters ontoereikend is. De beste schatting voor het aantal clusters wordt dan bereikt voor een geschat aantal singletonclusters dat sterk verschilt van het werkelijke aantal. Wanneer er voor verschillende collecties van documenten D met een constante $|D|$ wordt vastgesteld dat het geschat aantal clusters constant is, wil dit zeggen dat a overschat is.

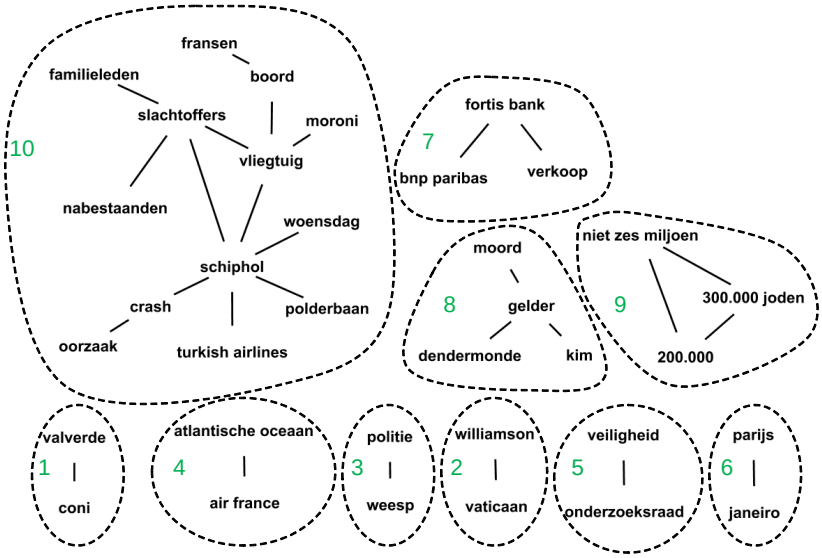
8.6 Clusteren van documenten

8.6.1 Basismethode

In Sectie 8.4 is aangetoond dat het expliciete gebruik van een evaluator voor documenten wordt bemoeilijkt door het onbekend zijn van de transformatiefunctie $\mathcal{T}$. Om dit probleem op te lossen zijn enkele veronderstellingen gemaakt over enerzijds de evaluator $E_{\mathcal{D}}$ en anderzijds de functie $\mathcal{T}$. Als gevolg van deze veronderstellingen kunnen we het coreferentieprobleem voor documenten herschrijven in termen van een relatie $R^{(\mathcal{D})} \subset \mathcal{C}^2$ en de selectieoperator $\mathcal{D}_{(c_1,c_2),\in}$. Er is ook verondersteld dat elk document d precies één entiteit beschrijft, zodat we op basis van de relatie $R^{(\mathcal{D})}$ en de selectieoperator een partitie van D moeten vormen, waarbij partitieklassen overeenstemmen met clusters.

In deze sectie bestuderen we hoe de onbekende relatie $R^{(\mathcal{D})}$ voor een gegeven collectie van documenten D kan worden bepaald. Hiervoor tonen we eerst aan hoe een partitie van D kan worden afgeleid op basis van een gegeven relatie $R^{(\mathcal{D})}$. Bij het maken van deze partitie voeren we een controle uit op de *zuiverheid* van partitieklassen. Dit betekent dat we een partitie afleiden zodat documenten in een partitieklasse met grote zekerheid hetzelfde onderwerp beschrijven. Daarna maken we gebruik van de schatting van het aantal clusters (Sectie 8.5) om een keuze te maken uit verschillende relaties $R^{(\mathcal{D})}$. Meer bepaald kiezen we een relatie $R^{(\mathcal{D})}$ die aanleiding geeft tot een partitie waarbij het aantal partitieklassen zo dicht mogelijk bij het geschat aantal clusters ligt. Onze aanpak kan bijgevolg worden samengevat in drie stappen:

- Zoek een relatie $R^{(\mathcal{D})} \subset \mathcal{C}^2$



Figuur 8.9: 12-snede van koppels van concepten

- Vorm clusters met hoge *zuiverheid* op basis van $R^{(\mathcal{D})}$
- Controleer het aantal clusters

Het zoeken naar een relatie $R^{(\mathcal{D})}$ betekent dat we koppels van relevante concepten moeten zoeken in het geheel van koppels. Door gebruik te maken van het relationele model voor tekst, kunnen we alle koppels in een collectie van documenten D bijzonder eenvoudig voorstellen als volgt:

$$\psi(D) = \bigoplus_{d \in D} \psi(d). \tag{8.46}$$

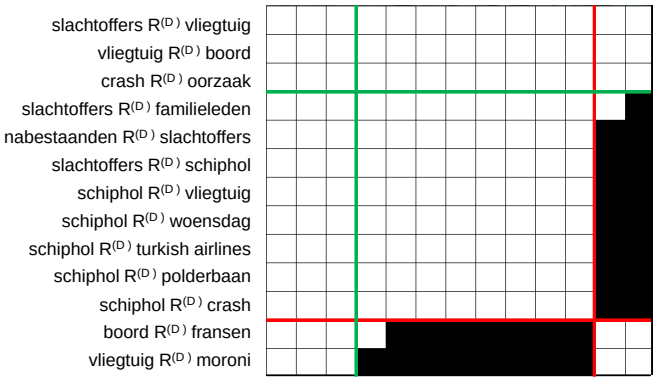
$\psi(D)$ is bijgevolg de multirelatie van alle koppels van concepten. Voor een koppel van concepten $(c_1, c_2) \in \mathcal{C}^2$ geeft het getal $\omega_{\psi(D)}(c_1, c_2)$ aan hoeveel keer dit koppel in totaal voorkomt in de relationele transformaties van alle documenten in D.

Laat ons bekijken wat er gebeurt wanneer we de multirelatie van koppels $\psi(D)$ afsnijden. Voor de verzameling van documenten die in Sectie 8.7 bij de experimenten wordt gebruikt, toont Figuur 8.9 de 12-snede. Deze snede is een binaire relatie en kan dus als graaf worden voorgesteld. De snede bestaat duidelijk uit verschillende componenten (genummerd en omlijnd in Figuur 8.9) in de graaf-theoretische zin van het woord. Dit wil zeggen dat de relatie verkregen als resultaat van een afsnijding, bestaat uit een aantal deelrelaties die onderling disjunct zijn. We willen in deze sectie komen tot een aanpak waarbij deze componenten worden gebruikt als basis voor de clustervorming. Bij een

dergelijke aanpak moeten we rekening houden met twee aspecten: afhankelijkheden binnen een component en afhankelijkheden tussen componenten. In wat volgt bespreken we eerst deze beide aspecten afzonderlijk en we bespreken het verband met de ingevoerde operatoren. Vervolgens geven we een algoritme dat, rekening houdend met afhankelijkheden, clusters van documenten vormt op basis van één snede. Ten slotte bespreken we hoe we, gegeven een collectie van documenten D, kunnen zoeken naar een optimale snede voor het vormen van clusters. Deze optimale snede wordt dan gebruikt voor de effectieve generatie van clusters.

Afhankelijkheden binnen een component Wanneer we elke component van een snede gebruiken voor de generatie van een cluster, moeten we rekening houden met het feit dat de patronen van een component niet noodzakelijk één onderwerp beschrijven. Beschouw bijvoorbeeld component 10 in Figuur 8.9. Deze component bevat zowel concepten die verwijzen naar een vliegtuigongeluk in Schiphol als concepten die verwijzen naar een vliegtuigongeluk in Moroni. We kunnen echter veronderstellen dat sommige patronen wel degelijk specifiek één onderwerp beschrijven. Met betrekking tot component 10 in Figuur 8.9 kunnen we bijvoorbeeld met grote zekerheid stellen dat het patroon "vliegtuig $R^{(\mathcal{D})}$ moroni" enkel zal voorkomen in documenten met als onderwerp het vliegtuigongeluk in Moroni. Dit komt voort uit de zeer specifieke, plaatsgebonden informatie gegeven door de plaatsnaam "Moroni". Ook kunnen we met grote zekerheid stellen dat het patroon "crash $R^{(\mathcal{D})}$ schiphol" enkel zal voorkomen in documenten met als onderwerp het vliegtuigongeluk in Schiphol. Ook hier wordt dit veroorzaakt door plaatsgebonden informatie. Om na te gaan of deze veronderstellingen kloppen, controleren we of de beide patronen onafhankelijk zijn. Immers als beide patronen onafhankelijk zijn, wil dat zeggen dat er geen enkel document bestaat waarin beide patronen samen voorkomen. Het in kaart brengen van onafhankelijkheden kan worden gedaan met behulp van de operatoren geïntroduceerd in Sectie 8.3. Meer bepaald kunnen alle afhankelijkheden binnen een component in kaart worden gebracht door analyse van de matrix $\mathbf{M}_{\mathbf{v}^*}$, waarbij $\mathbf{v}$ de vector van patronen in de component voorstelt. Voor component 10 van de 12-snede toont Figuur 8.10 de matrix $\mathbf{M}_{\mathbf{v}^*}$.

Het bestuderen van deze matrix maakt duidelijk dat er twee onderwerpen worden vertegenwoordigd door de component. Deze onderwerpen manifesteren zich als blokken van afhankelijke patronen op de diagonaal van de matrix $\mathbf{M}_{\mathbf{v}^*}$. De vliegtuigramp op de luchthaven van Schiphol wordt vertegenwoordigd door de diagonaalblok afgelijnd door de twee groene en de twee rode lijnen. De vliegtuigramp in Moroni wordt vertegenwoordigd door het 2×2-blok onderaan rechts. Er zijn een aantal patronen in de component aanwezig die geen onderscheid kunnen maken tussen de beide vliegtuigrampen, aangezien er voor beide onderwerpen documenten bestaan waarin deze patronen voorkomen. Deze patronen komen overeen in Figuur 8.10 met het 3×3-blok bovenaan links. Hieruit kunnen we afleiden dat, als voor een component van de snede, met patronen



Figuur 8.10: Afhankelijkheden binnen een snedecomponent

$\mathbf{v}$, een cluster wordt gegenereerd als volgt:

$$\bigcup_{(c_1,c_2)\in\mathbf{v}} D_{(c_1,c_2),\in} \tag{8.47}$$

dan laat $\mathbf{M}_{\mathbf{v}^*}$ toe om de *zuiverheid* van deze cluster te controleren door de onafhankelijkheden in de component in kaart te brengen. Deze *zuiverheid* kan worden geoptimaliseerd door de patronen van een component op te splitsen in verschillende collecties van patronen.

Algoritme 8.2 split($\mathbf{M}_{\mathbf{v}^*}$)

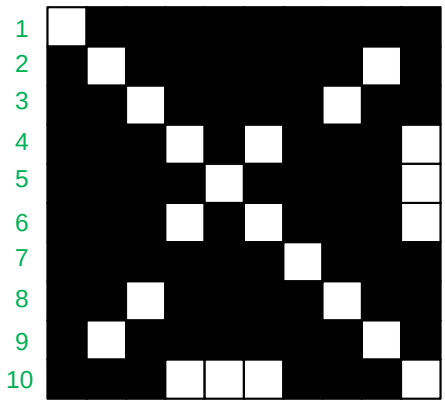
Gegeven: splitsfactor $sf \in [0,1]$

1: **for** $j \in \{1,...,|\mathbf{v}^*|\}$ **do**
2: $\quad x \leftarrow \sum_{k=1}^{|\mathbf{v}^*|}(\mathbf{M}_{\mathbf{v}^*}(j,k))$
3: $\quad$ **if** $\frac{x}{|\mathbf{v}^*|} < sf$ **then**
4: $\quad\quad$ **return** $\{\mathbf{v}^*_{(1,...,j)}\} \cup \{\mathbf{split}(\mathbf{M}_{\mathbf{v}^*_{(j+1,...,|\mathbf{v}^*|)}})\}$
5: $\quad$ **end if**
6: **end for**

Algoritme 8.2 toont de pseudocode voor een dergelijke opsplitsing. De matrix $\mathbf{M}_{\mathbf{v}^*}$ wordt doorlopen tot het aantal onafhankelijke patronen te groot wordt (regel 3). De patronen die tot dan geobserveerd zijn, worden samen afgesplitst en de rest van de matrix wordt recursief onderzocht (regel 4). Voor de component uit Figuur 8.10 levert dit twee collecties van patronen, waarbij de kleinste collectie bestaat uit de patronen ("vliegtuig $R^{(\mathcal{D})}$ moroni") en ("boord $R^{(\mathcal{D})}$ fransen").

Afhankelijkheden tussen componenten Naast de afhankelijkheden binnen componenten, zijn ook afhankelijkheden tussen componenten belangrijk. Meer bepaald, als een strategie gevolgd wordt waarbij clusters voortkomen uit

één component, dan is die strategie foutgevoelig op het gebied van de *compleetheid* van clusters. Het kan bijvoorbeeld zijn dat patronen van verschillende componenten eenzelfde onderwerp beschrijven. Wanneer dit zo is, dan moeten er documenten zijn die patronen uit verschillende componenten bevatten. Bijgevolg moeten er dan afhankelijkheden tussen componenten bestaan.



Figuur 8.11: Afhankelijkheden tussen snedecomponenten

Figuur 8.11 toont voor de snede uit Figuur 8.9, een matrix van afhankelijkheden tussen componenten, waarbij afhankelijkheid tussen componenten een rechtstreekse uitbreiding is van afhankelijkheid tussen patronen. Meer bepaald, als we een component noteren als een vector $\mathbf{v}$ van patronen, dan zijn twee componenten $\mathbf{v}$ en $\mathbf{v}'$ afhankelijk, genoteerd als $\mathbf{v} \sim \mathbf{v}'$, als:

$$\exists i \in \{1, ..., |\mathbf{v}|\} : \exists j \in \{1, ..., |\mathbf{v}'|\} : \mathbf{v}_i \sim \mathbf{v}'_j \tag{8.48}$$

Op een gelijkaardige manier kunnen we ook het concept 'afhankelijkheidsgraad' uitbreiden naar componenten als volgt:

$$\text{dep}(\mathbf{v}, \mathbf{v}') = \frac{\left|\left(\bigcup_{i=1}^{|\mathbf{v}|} \mathcal{D}_{\mathbf{v}_i,\in}\right) \cap \left(\bigcup_{i=1}^{|\mathbf{v}'|} \mathcal{D}_{\mathbf{v}'_i,\in}\right)\right|}{\min\left(\left|\bigcup_{i=1}^{|\mathbf{v}|} \mathcal{D}_{\mathbf{v}_i,\in}\right|, \left|\bigcup_{i=1}^{|\mathbf{v}'|} \mathcal{D}_{\mathbf{v}'_i,\in}\right|\right)}. \tag{8.49}$$

Figuur 8.11 toont dat tussen componenten wel degelijk afhankelijkheden bestaan die in rekening gebracht moeten worden. In de strategie die hier wordt nagestreefd, zullen componenten die zeer sterk afhankelijk zijn met een grotere component (in termen van aantal patronen) worden verwijderd. Sterk afhankelijk dient hierbij geïnterpreteerd te worden als een hoge afhankelijkheidsgraad. Als deze afhankelijkheidsgraad tussen twee componenten hoog genoeg is (d.i. groter dan een drempelwaarde), kan met grote zekerheid worden gezegd dat de beide componenten handelen over hetzelfde onderwerp. De keuze om de component met het kleinst aantal patronen te verwijderen eerder dan beide componenten samen te voegen, wordt gemotiveerd door drie argumenten. Ten

eerste impliceert een hoge afhankelijkheidsgraad per definitie dat beide componenten bestaan uit patronen waarvoor de relationele selecties sterk overlappen. Aangezien beide componenten een bepaald aspect beschrijven van het onderwerp in kwestie, betekent de hoge afhankelijkheidsgraad dat de informatie die beide componenten bieden, sterk overlapt. Er zal dus weinig verlies in *compleetheid* optreden als de kleinere component wordt genegeerd. Ten tweede levert het verwijderen van de kleinste component een voordeel in complexiteit, door het kleiner aantal afhankelijkheden dat nadien moet worden geanalyseerd. Het stelselmatig splitsen van een component met Algoritme 8.2 heeft immers een kwadratische complexiteit in termen van het aantal patronen in de component. Ten derde leren experimenten dat de aanwezigheid van veel componenten met slechts één patroon kan leiden tot foutpropagatie. Meer bepaald, door het samenvoegen van twee componenten met elk één patroon ontstaan sterkere afhankelijkheden tussen componenten, die kunnen leiden tot foute samenvoeging van componenten.

Clustervorming op basis van een snede Gelet op de (on)afhankelijkheden binnen componenten en de (on)afhankelijkheden tussen componenten, geven we nu een algoritme om op basis van een k-snede van $\psi(D)$ een partitie van D te vormen. Deze methode wordt in pseudocode samengevat in Algoritme 8.3.

Algoritme 8.3 produce(D,k)

Gegeven: verwijderfactor $df \in [0,1]$

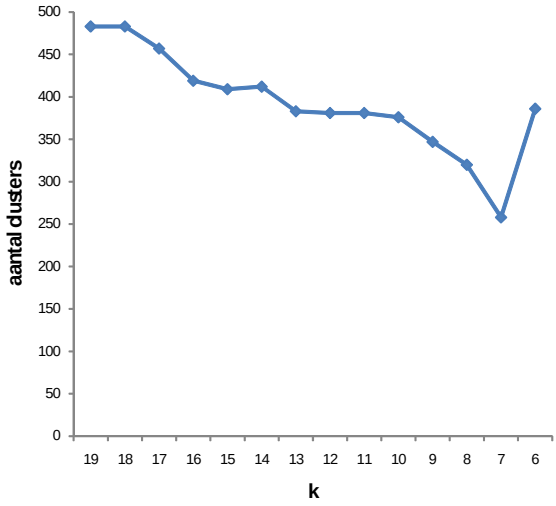
1: $K \leftarrow \emptyset$
2: $R_k^{(\mathcal{D})} \leftarrow \psi(D)_k$
3: $\mathcal{R}_k \leftarrow$ **components**$(R_k^{(\mathcal{D})})$
4: **while** $\mathcal{R}_k \neq \emptyset$ **do**
5: $\quad \mathbf{v} \leftarrow \arg\min_{\mathbf{v} \in \mathcal{R}_k} |\mathbf{v}|$
6: $\quad$ **if** $\neg (\exists \mathbf{v}' \in \mathcal{R}_k : \mathrm{dep}(\mathbf{v}, \mathbf{v}') > df)$ **then**
7: $\quad\quad (\mathbf{v}^1, ..., \mathbf{v}^l) \leftarrow$ **split**$(\mathbf{M}_{\mathbf{v}^*})$
8: $\quad\quad$ **for** $j \in \{l, ..., 1\}$ **do**
9: $\quad\quad\quad cluster \leftarrow \bigcup_{(c_1,c_2) \in \mathbf{v}^j} D_{(c_1,c_2),\in}$
10: $\quad\quad\quad$ **add**$(K, cluster)$
11: $\quad\quad\quad D \leftarrow D \backslash cluster$
12: $\quad\quad$ **end for**
13: $\quad$ **end if**
14: $\quad \mathcal{R}_k \leftarrow \mathcal{R}_k \backslash \mathbf{v}$
15: **end while**
16: **return** K

Algoritme 8.3 maakt een snede van de relationele documentrepresentatie van D (regel 2). Deze snede wordt genoteerd als $R_k^{(\mathcal{D})}$. Vervolgens wordt deze snede opgesplitst in maximaal geconnecteerde componenten (regel 3). De verzameling van deze componenten wordt genoteerd als $\mathcal{R}_k$. Deze opsplitsing wordt hier niet beschreven, aangezien het een basisprobleem uit de grafenthe-

orie betreft en de oplossing van dit probleem een lineaire complexiteit heeft [129]. De componenten worden gesorteerd volgens grootte (d.i. aantal patronen). De kleinste componenten krijgen voorrang aangezien zij onderwerpen vertegenwoordigen waarvoor het minste informatie voorhanden is. Deze onderwerpen zijn bijgevolg ook het moeilijkst te identificeren. Vervolgens wordt gecontroleerd of voor de geselecteerde component een grotere component bestaat, waarmee de geselecteerde component een sterke afhankelijkheid vertoont, d.i. een afhankelijkheidsgraad heeft die groter is dan een gegeven drempelwaarde *df* (regel 6). Wanneer geen sterke afhankelijkheden bestaan wordt elke component opgesplitst in deelcomponenten (regel 7). De lijst van deelcomponenten wordt dan van achter naar voren doorlopen. De volgorde zorgt ervoor dat de kleinste deelcomponenten opnieuw eerst worden gekozen. Het is namelijk een eigenschap van $\mathbf{M}_{\mathbf{v}^*}$ dat de grootste diagonaalblokken zich steeds linksboven in de matrix bevinden, waardoor de grootste deelcomponenten eerst worden afgesplitst. Voor elke deelcomponent wordt een cluster gemaakt die bestaat uit de unie van de relationele selecties (Definitie 8.8) van de patronen in de deelcomponent. De documenten uit deze cluster worden vervolgens uit D verwijderd, zodat de clusters onderling disjunct zijn (d.i. zodat de clusters overeenkomen met partitieklassen).

Bepaling van de optimale snede We hebben nu een algoritme dat voor een collectie van documenten D, voor elke k-snede van $\psi(D)$, een partitie van D genereert. Hierbij wordt $(\psi(D))_k$ gezien als een mogelijke kandidaat voor de gezochte relatie $R^{(\mathcal{D})}$. Er blijft nu nog de vraag hoe de waarde van k moet worden bepaald. Door gebruik van Algoritme 8.2 worden patronen binnen één component gegroepeerd zodat de unie van de relationele selecties van patronen in één groep aanleiding geeft tot een cluster met een voldoende hoge *zuiverheid*. Het kan enerzijds worden ingezien dat de betrouwbaarheid van de schatting van *zuiverheid* lager is wanneer componenten zeer weinig patronen bevatten. Anderzijds, wanneer componenten zeer veel patronen bevatten, wil dit typisch zeggen dat ze deel uitmaken van een snede met lage multipliciteit, hetgeen per definitie wil zeggen dat veel patronen onafhankelijk zullen zijn van andere patronen. Een teveel aan patronen leidt tot een verspreiding van coreferente documenten over verschillende clusters. Dit betekent dat de *compleetheid* van clusters laag zal zijn. Er moet dus een snede worden gevonden waarin de balans tussen *zuiverheid* en *compleetheid* wordt geoptimaliseerd.

Deze optimale snede kan worden gevonden als volgt. Wanneer Algoritme 8.3 wordt uitgevoerd voor verschillende waarden van k, kan telkens worden opgemeten hoeveel clusters worden geproduceerd. Hierbij moeten documenten uit D die niet geselecteerd zijn, als een singletoncluster worden beschouwd. Deze metingen kunnen worden vergeleken met het resultaat van de schatting van het aantal clusters (Sectie 8.5). Figuur 8.12 toont de evolutie van het aantal gegenereerde clusters voor verschillende snedes. Aangezien de schatting van het aantal clusters voor deze documenten 137 bedraagt, is de optimale snede deze met $k = 7$. De reden dat voor $k = 6$ het aantal clusters weer stijgt, is



Figuur 8.12: Aantal clusters voor verschillende k-snedes

dat componenten van deze snede een teveel aan weinig voorkomende patronen bevatten, waardoor versplintering optreedt. Dit betekent dat heel veel clusters gegenereerd worden waarbij elke cluster weinig documenten bevat.

Dit geeft aanleiding tot een strategie waarbij voor verschillende snedes clusters worden geproduceerd en waarbij vervolgens die snede wordt gekozen, waarvoor het verkregen aantal clusters het dichtst bij het geschatte aantal clusters ligt. Het gebruik van een dergelijke strategie zorgt tegelijk voor een controle van de *zuiverheid* (door controle van de afhankelijkheden binnen componenten) en controle van de *compleetheid* (schatting van het aantal clusters). Na het vinden van de optimale snede en het produceren van de eerste clusters, kan deze strategie worden herhaald voor de documenten die onder de optimale snede niet zijn geselecteerd. Dit is nodig wanneer het aantal clusters voor de optimale snede nog boven het geschatte aantal clusters ligt.

Algoritme 8.4 vat de basis van onze methode als volgt samen. Gegeven een verzameling van documenten, schat het aantal entiteiten dat aanwezig is (regel 1). Zoek vervolgens een snede waarvoor het geproduceerde aantal clusters zo dicht mogelijk bij deze schatting ligt (regel 5). Produceer een reeks van clusters met deze optimale snede (regel 7). Wanneer het bekomen aantal clusters boven het geschatte aantal clusters ligt, zoek dan voor de overgebleven documenten opnieuw een optimale snede (regel 9). Produceer clusters met deze nieuwe snede en voeg deze toe aan de reeds bekomen clusters. Deze procedure herhaalt zich tot het gewenste aantal clusters bereikt is of tot de optimale snede het resultaat is van een te lage k (regel 6). Algoritme 8.4 maakt voor de controle van dit laatste criterium gebruik van een snede factor cf.

De basismethode beschouwt drie parameters: de splitsfactor sf, de verwij-

Algoritme 8.4 cluster(D)

Gegeven: snedefactor $cf \in \mathbb{N}$

$b \leftarrow$ **estimate**(D)
$stop \leftarrow$ **false**
$K \leftarrow \emptyset$
while $\neg stop$ **do**
  $k_{optimal} \leftarrow \arg\min_k \left|\left(|\mathbf{produce}(D,k)| + |D \backslash D_{\mathbf{produce}(D,k)}| - b\right)\right|$
  if $k_{optimal} > cf$ **then**
    $K \leftarrow K \cup \mathbf{produce}(D, k_{optimal})$
    $D \leftarrow D \backslash D_K$
    $stop \leftarrow |K| + |D| < b$
  else
    $stop \leftarrow$ **true**
  end if
end while
$\forall d \in D : K \leftarrow K \cup \{d\}$
return K

derfactor df en de snedefactor cf. De splitsfactor sf laat toe om te bepalen wanneer componenten moeten worden opgesplitst in deelcomponenten, door een grens te leggen op het aantal onafhankelijke patronen in de component. Deze factor moet enerzijds hoog genoeg worden gekozen om de *zuiverheid* te garanderen, maar anderzijds mag een klein aantal onafhankelijkheden niet leiden tot een splitsing. In experimenten wordt voor de splitsfactor typisch de waarde 0.5 gekozen, om deze afweging in kaart te brengen. De verwijderfactor df, die in kaart brengt wanneer componenten sterk afhankelijk zijn van elkaar, moet per definitie hoog worden gekozen. Zoniet gaat de notie 'sterk afhankelijk' verloren. Typisch wordt deze factor niet lager dan 0.8 gekozen. De snedefactor cf ten slotte, is typisch sterk invariant over verschillende problemen heen. Deze factor moet verhinderen dat het merendeel van de patronen wordt gebruikt om clusters te produceren. Dit wil zeggen dat voor cf typisch de waarde 2 of 3 gekozen wordt. Een keuze van 2 of minder zal leiden tot een groot aantal patronen dat in kaart wordt gebracht. Dit is nefast voor de accuraatheid van de methode en impliceert bovendien een bijzonder hoge rekentijd door het opblazen van het aantal patronen per component. Een keuze van meer dan 3 zou dan weer leiden tot het niet in beschouwing nemen van relevante patronen. De conclusie voor elk van de drie factoren is dat zij typisch maar een beperkt aantal zinvolle waarden hebben.

8.6.2 Uitbreidingen van het algoritme

In deze sectie bespreken we twee uitbreidingen van de basismethode die tot hier toe is opgebouwd.

Alternatieve evaluatoren Ten eerste is in Sectie 8.4 aangehaald dat de keuze voor een andere selectieoperator alternatieve evaluatoren definieert. Deze evaluatoren geven aanleiding tot selectie van documenten via de conceptuele selectie (Definitie 8.9):

$$\mathcal{D}_{(c_1,c_2),\,\widehat{\in}} \tag{8.50}$$

en de possibilistische selectie (Definitie 8.10)

$$\mathcal{D}_{(c_1,c_2),\,\widetilde{\in}}\,. \tag{8.51}$$

Op basis van Stelling 8.1 weten we dat evaluatoren kunnen worden geordend, hetgeen overeenkomt met een ordening van de overeenkomstige selectieoperatoren. De alternatieve evaluatoren geven aanleiding tot grotere clusters, waardoor de *zuiverheid* daalt en de *compleetheid* stijgt. In het geval van $E_{\mathcal{D}}^{\widetilde{\in}}$ zullen we concepten vergelijken met een evaluator $E_{\mathcal{S}}^{*}$ met een parameterkeuze $(1, 0, 0.05)$ die voor korte strings typisch goed werkt (Hoofdstuk 6). Het gebruik van een evaluator $E_{\mathcal{S}}^{*}$ veroorzaakt uiteraard een bijkomende complexiteit in het berekenen van clusters. De ervaring leert dat het gebruik van alternatieve evaluatoren vooral interessant is wanneer een component weinig patronen bevat. In een praktische situatie blijkt een goede afweging tussen complexiteit en accuraatheid te worden gevonden wanneer $\mathcal{D}_{(c_1,c_2),\,\widetilde{\in}}$ enkel wordt gebruikt voor componenten met slechts één patroon. Grotere componenten zijn namelijk zonder het gebruik van $\mathcal{D}_{(c_1,c_2),\,\widetilde{\in}}$ in staat de meerderheid van de gezochte documenten te vinden, daar waar dit voor componenten met slechts één patroon niet altijd het geval is.

Samenvoeging van clusters Ten tweede kan de basismethode worden uitgebreid door koppels van clusters te zoeken waarvoor samenvoeging de *compleetheid* verhoogt. Dit kan niet tijdens de uitvoering van het algoritme, aangezien pas na splitsing wordt verondersteld dat componenten aanleiding geven tot clusters met een hoge *zuiverheid*. Sterke afhankelijkheden tussen componenten duiden enerzijds op eenzelfde onderwerp, maar anderzijds kan één component overeenkomen met verschillende onderwerpen. Twee componenten samenvoegen is daarom geen geschikte oplossing, zoals ook is uitgelegd bij het beschrijven van de controle op sterk afhankelijke componenten. Pas na het uitvoeren van het clusteralgoritme verkrijgen we clusters waarvoor de *zuiverheid* is geverifieerd. Dit is het geschikte punt om te controleren of clusters alsnog kunnen worden samengevoegd. Deze samenvoeging gebeurt als volgt. We beschouwen twee clusters $X \subset D$ en $X' \subset D$ en de bijhorende patronen $\mathbf{v}$ en $\mathbf{v}'$. Veronderstel nu de vector van patronen:

$$\mathbf{v}'' = \left(\mathbf{v}_1, ..., \mathbf{v}_{|\mathbf{v}|}, \mathbf{v}'_1, ..., \mathbf{v}'_{|\mathbf{v}'|}\right). \tag{8.52}$$

We kunnen dan op basis van $\mathbf{M}_{\mathbf{v}''}$ de afhankelijkheden tussen patronen uit $\mathbf{v}$ en $\mathbf{v}'$ bestuderen. We stellen hier de aanpak voorop dat X en X' samengevoegd mogen worden als er voldoende afhankelijkheden bestaan tussen de patronen

$\mathbf{v}$ en $\mathbf{v}'$. Anders gezegd, X en X' mogen worden samengevoegd als:

$$\sum_{i=1}^{|\mathbf{v}''|}\sum_{i=1}^{|\mathbf{v}''|}\mathbf{M}_{\mathbf{v}''}(i,j) \tag{8.53}$$

boven een gegeven drempelwaarde ligt.

Een probleem stelt zich wanneer zowel $\mathbf{v}$ als $\mathbf{v}'$ een heel klein aantal patronen bevat. Een analyse van afhankelijkheden biedt dan weinig inzicht in het coreferent zijn van de documenten uit de overeenkomende clusters. Hoe groter het aantal patronen dat in één van de twee clusters aangetroffen wordt, hoe accurater de besluittrekking voor samenvoeging. Om die reden kunnen clusters pas worden samengevoegd als één van beide clusters voldoende patronen bevat. In onze aanpak stellen we daarom voorop dat:

$$\max\left(|\mathbf{v}|, |\mathbf{v}'|\right) > 1. \tag{8.54}$$

8.7 Experimenten

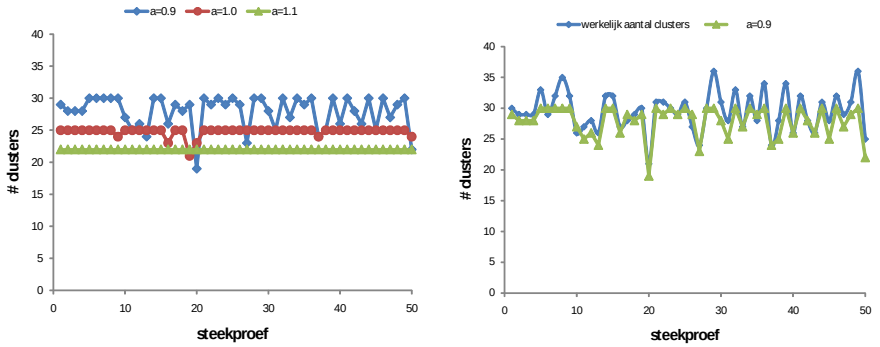
In deze sectie worden enkele experimenten beschreven waarbij de prestaties van de geïntroduceerde methode voor tekstclustering worden bestudeerd. Eerst wordt onderzocht hoe goed het aantal clusters kan worden geschat. Daarna wordt bekeken hoe de aanpak voor clustering zich verhoudt tot bestaande methoden.

8.7.1 Schatting van het aantal clusters

Eerst en vooral zal met een experiment worden aangetoond hoe de schatting van het aantal clusters presteert. Dit experiment verloopt in twee stappen. In een eerste stap wordt het aantal singletonclusters als gekend verondersteld. Op basis van dit aantal wordt dan een schatting gemaakt van het aantal clusters. Dit experiment moet aantonen dat een schatting van het aantal clusters, gebaseerd op het aantal singletonclusters, wel degelijk mogelijk is. In een tweede stap is het aantal singletonclusters niet gekend en moet het worden benaderd aan de hand van relevante concepten.

In eerste instantie veronderstellen we dus dat het aantal singletonclusters gekend is. Beschouwen we de eerder geïntroduceerde datacollectie van 550 documenten met daarin 133 onderwerpen, dan kunnen we steekproeven trekken van verschillende groottes. Zoals aangehaald in Sectie 8.5 veronderstellen we dat de toekenning van documenten aan onderwerpen voldoet aan de Wet van Zipf, waarbij twee parameters a en b worden beschouwd. Parameter a is een vormparameter en dient hierbij in een nauw interval te worden gekozen. Parameter b stelt het aantal onderwerpen voor.

Laat ons voor een herhaalde steekproef van 50 documenten bekijken hoe de schatting van het aantal clusters varieert in functie van variaties van de vormparameter a. Figuur 8.13 leert ons dat wanneer parameter a te groot wordt



Figuur 8.13: Bepaling van $|\,\widehat{(\mathcal{E}_D)_1}\,|$ voor verschillende a (links) en vergelijking met $|\,(\mathcal{E}_D)_1\,|$ (rechts)

gekozen, er geen of nauwelijks variatie is in de schatting. Voor verschillende steekproeven van 50 documenten blijft de schatting nagenoeg constant. Door een te hoge waarde voor a kan geen goede verdeling worden gevonden. Wanneer we voor a echter de grootste waarde zoeken waarvoor de variatie in de schatting significant groter is dan 0, blijkt dit een geschikte waarde te zijn. In Figuur 8.13 wordt, voor de steekproeven van 50 documenten onder deze keuze voor a, een vergelijking getoond met het effectieve aantal clusters. Deze figuur toont een vrij nauwkeurige schatting van het aantal clusters, waaruit blijkt dat a relatief eenvoudig kan worden gevonden.

Om uitdrukking te geven aan de gemaakte fout, beschouwen we de absolute gemiddelde fout voor n schattingen als volgt:

$$AME = \sum_{i=1}^{n} \left| (|\,\widehat{(\mathcal{E}_D)_1}\,|) - (|\,(\mathcal{E}_D)_1\,|) \right|. \tag{8.55}$$

Tabel 8.1 toont de AME voor steekproeven van verschillende grootte. De vierde kolom toont de AME in het geval waar het aantal singletonclusters gekend is. Deze evolutie van de AME in functie van de steekproefgrootte laat zien dat de schatting van het aantal clusters betrouwbaar is in vergelijking met het gemiddeld aantal clusters voor de overeenkomstige steekproefgrootte. Bijvoorbeeld, voor 300 documenten zijn gemiddeld 91 clusters aanwezig. Indien het aantal singletonclusters gekend is, zal de schatting van het aantal clusters een gemiddelde absolute fout van 2.26 vertonen, wat ongeveer een fout van 2.5% betekent. Uiteraard illustreren deze cijfers enkel dat de idee achter singletonclusters werkt. In realiteit is het aantal singletonclusters niet gekend en moet dit aantal worden geschat. Indien dit gebeurt, veroorzaakt dit een bijkomende fout in de schatting van het aantal clusters. De impact van deze bijkomende fout wordt getoond in de vijfde kolom van Tabel 8.1, waaruit blijkt dat de gemiddelde absolute fout op de schatting van het aantal clusters ongeveer verdubbelt. De kwaliteit van onze methode kan worden afgeleid uit deze cijfers.

Hoewel de foutpercentages soms hoog kunnen lijken, zal blijken uit de experimenten verder in dit hoofdstuk dat de accuraatheid volstaat voor het doel dat we hier voor ogen hebben, namelijk het zoeken naar coreferente documenten.

$\|D\|$	gem. $\|(\mathcal{E}_D)_1\|$	a	AME σ gekend (%)	AME σ geschat (%)
50	29	0.90	1.40 (4.83%)	2.16 (7.45%)
100	46	0.90	1.76 (3.83%)	3.94 (8.57%)
150	59	1.00	1.98 (3.36%)	3.46 (5.86%)
200	70	1.00	1.36 (1.94%)	4.80 (6.86%)
250	81	1.05	1.66 (2.05%)	3.88 (4.79%)
300	91	1.05	2.26 (2.50%)	4.10 (4.51%)

Tabel 8.1: Resultaten voor bepaling van het aantal clusters

8.7.2 Clusteren van documenten

In deze sectie wordt onze aanpak voor clusteren van documenten bestudeerd aan de hand van vergelijkende experimenten. Hiervoor worden enkele methoden uit de literatuur toegepast op de volledige datacollectie beschreven in Sectie 8.1 en wordt een vergelijking gemaakt met onze aanpak. We bestuderen de verschillende varianten voor selectie van documenten op basis van patronen. Ten slotte willen we de robuustheid van onze methode onderzoeken door verschillende steekproeven uit de datacollectie te halen, waardoor telkens andere datacollecties worden verkregen.

Om te beginnen wordt uitgelegd hoe de resultaten worden gerapporteerd. In elk experiment wordt vertrokken van een verzameling documenten D. Na uitvoering van een clusteralgoritme wordt een verzameling van clusters verkregen. Deze clusters zijn een partitie van de verzameling D. Voor elke combinatie van onderwerp e_i en cluster K_j kunnen de *zuiverheid* $\text{zuiv}(e_i, K_j)$ en de *compleetheid* $\text{comp}(e_i, K_j)$ als volgt worden berekend:

$$\text{zuiv}(e_i, K_j) = \frac{n_{i,j}}{|K_j|} \tag{8.56}$$

$$\text{comp}(e_i, K_j) = \frac{n_{i,j}}{\omega_{\mathcal{E}_D}(e_i)} \tag{8.57}$$

waarbij $n_{i,j}$ gelijk is aan het aantal documenten over onderwerp e_i in K_j. Een afweging van beide metingen is de f-waarde, die voor een gegeven onderwerp e_i en cluster K_j wordt berekend als:

$$f(e_i, K_j) = \frac{2\ \text{zuiv}(e_i, K_j)\ \text{comp}(e_i, K_j)}{\text{zuiv}(e_i, K_j) + \text{comp}(e_i, K_j)}. \tag{8.58}$$

De accuraatheid van een clusteralgoritme kan dan worden uitgedrukt als een lineaire combinatie van maximale f-waarden voor de individuele clusters:

$$f = \sum_i \frac{\omega_{\mathcal{E}_D}(e_i)}{|D|} \max_j f(e_i, K_j). \tag{8.59}$$

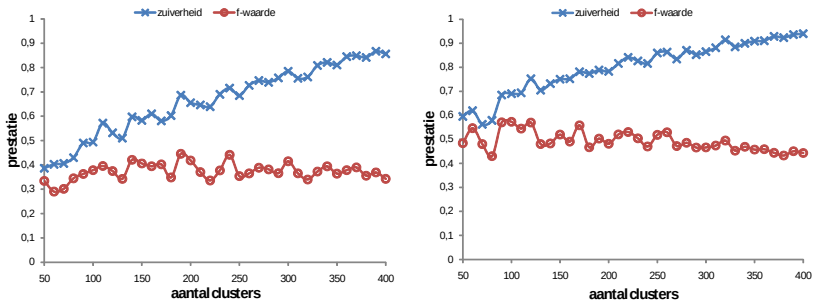
Om meer inzicht te krijgen in de samenstelling van de clusters wordt ook de globale *zuiverheid* van de clusters berekend als:

$$\text{zuiv} = \sum_j \frac{|K_j|}{|D|} \max_i \text{zuiv}(e_i, K_j). \tag{8.60}$$

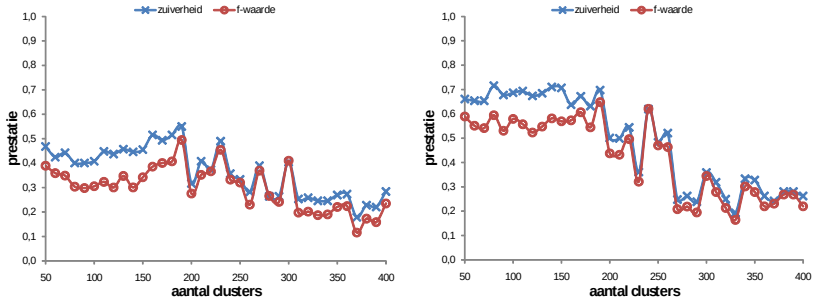
De selectie van de methoden uit de literatuur is gebaseerd op een recente studie die een overzicht geeft van methoden voor tekstclustering [127]. Er wordt voor deze methoden een vectorruimte vooropgesteld. De dimensies van de vectoren in deze vectorruimte kunnen worden ingevuld door een TFIDF-gewicht (zie [89] en Hoofdstuk 6) of door een binaire indicator die aangeeft of een woord voorkomt in een document. Beide mogelijkheden worden onderzocht. De studie [127] stelt dat de voorverwerking van documenten essentieel is om een goede tekstclustering te bekomen. Volgende voorverwerkingsstappen worden daarom toegepast [127]:

- **Filtering**: Leestekens en speciale karakters worden verwijderd.
- **Afsplitsing**: Documenten worden opgesplitst in woorden (1-grammen).
- **Verwijdering van stopwoorden**: Een lijst van woorden wordt gebruikt om irrelevante woorden te verwijderen (lidwoorden, voorzetsels ...).
- **Stamafleiding**: Afgesplitste woorden worden gereduceerd tot een stamvorm. Meervouden, verkleinwoorden ... worden hierdoor teruggebracht tot een basisvorm [130].
- **Frequentiefiltering**: Woorden die over de hele datacollectie minder dan een gegeven drempelwaarde voorkomen, worden verwijderd uit de datacollectie.
- **PCA**: De dimensie van de vectoren wordt verder beperkt op basis van een singuliere waardendecompositie zodat 95% van de variantie wordt verklaard (Bijlage A).

De volledige datacollectie bestaat uit 14053 dimensies. Na filteren en stamafleiding blijven nog 4590 dimensies over. Deze dimensies worden vervolgens geanalyseerd met PCA. In het geval van binaire vectoren levert dit 383 dimensies op daar waar het geval van vectoren met TFIDF-gewicht leidt tot 372 dimensies. Op de resulterende matrices worden volgende methoden uit de literatuur toegepast: *k-means* clusteren, kernel *k-means* clusteren en hiërarchisch clusteren. Er wordt verwezen naar Bijlage A voor een introductie tot deze methoden. Aangezien voor elk van deze methoden het aantal clusters als gekend wordt beschouwd, wordt het effect van het aantal clusters op de *zuiverheid* en de f-waarde bestudeerd. Figuur 8.14 toont de evolutie van de *zuiverheid* en f-waarde voor *k-means* clusteren over verschillende waarden van $|\,(\mathcal{E}_D)_1\,|$ (d.i., het aantal clusters). Hierbij wordt vastgesteld dat het gebruik van TFIDF-vectoren een iets beter resultaat geeft dan het gebruik van binaire vectoren.



Figuur 8.14: *Zuiverheid* en f-waarde in functie van het aantal clusters $|\,(\mathcal{E}_D)_1\,|$ voor *k-means* clusteren: binaire vectoren (links) en TFIDF-vectoren (rechts)

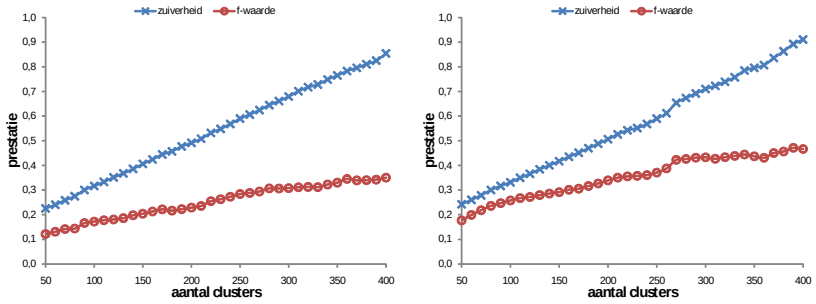


Figuur 8.15: *Zuiverheid* en f-waarde in functie van het aantal clusters $|\,(\mathcal{E}_D)_1\,|$ voor kernel *k-means* clusteren: binaire vectoren (links) en TFIDF-vectoren (rechts)

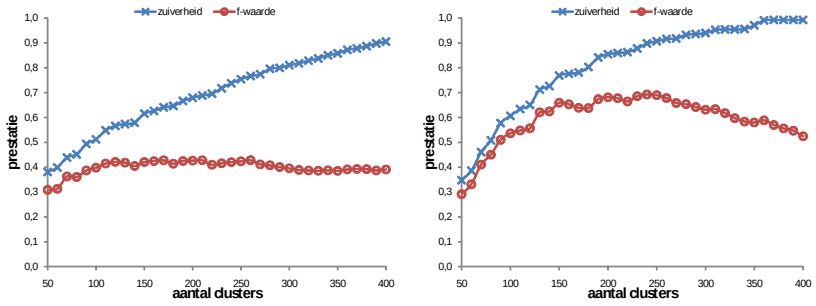
De stijgende trend in de *zuiverheid* is te wijten aan het steeds groter wordende aantal clusters. Hieruit besluiten we dat het *k-means* algoritme probeert om het opgegeven aantal clusters altijd te vullen. De quasi-constante trend in de f-waarde duidt op een dalende *compleetheid*, d.i. als de f-waarde gelijk blijft en de *zuiverheid* stijgt, dan moet de *compleetheid* dalen. Gelet op het feit dat het *k-means* algoritme enkel in staat is om lineair scheidbare clusters te genereren, wordt voor data met hoge dimensies vaak een niet-lineaire afbeelding naar een nieuwe ruimte gebruikt. Hiervoor wordt gebruik gemaakt van een kernelfunctie κ. In de context van documenten is gebleken dat de cosinus kernelfunctie een goed resultaat geeft [127]:

$$\kappa_{\cos}(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \, \|\mathbf{y}\|}. \tag{8.61}$$

De resultaten voor kernel *k-means* worden getoond in Figuur 8.15. We stellen vast dat er een opmerkelijk dalende trend is voor de *zuiverheid* in functie van stijgende $|\,(\mathcal{E}_D)_1\,|$. Dit is te wijten aan het feit dat kernel *k-means* heel wat lege clusters vormt. Deze lege clusters worden verwijderd tijdens de uitvoering van het algoritme. De dalende trend in de f-waarde toont aan dat dit niet accuraat



Figuur 8.16: *Zuiverheid* en f-waarde in functie van het aantal clusters $|\,(\mathcal{E}_D)_1\,|$ voor hiërarchisch clusteren (*enkelvoudig regel*): binaire vectoren (links) en TFIDF-vectoren (rechts)

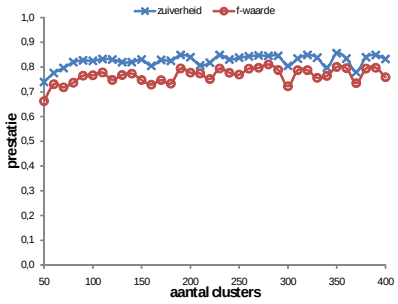


Figuur 8.17: *Zuiverheid* en f-waarde in functie van het aantal clusters $|\,(\mathcal{E}_D)_1\,|$ voor hiërarchisch clusteren (*volledige regel*): binaire vectoren (links) en TFIDF-vectoren (rechts)

gebeurt.

Een volgende methode uit de literatuur is het hiërarchische clusteralgoritme. Hierbij worden iteratief clusters samengevoegd. De regel voor het kiezen van de clusters die worden samengevoegd in een volgende iteratie, is een parameter van het algoritme. We beschouwen hier de *enkelvoudige regel* en de *volledige regel*. De *enkelvoudige regel* kiest die twee clusters waarvoor de minimale afstand tussen twee objecten uit beide clusters zo klein mogelijk is. De *volledige regel* kiest die twee clusters waarvoor de maximale afstand tussen twee objecten uit beide clusters zo klein mogelijk is. Figuren 8.16 en 8.17 tonen de resultaten voor beide regels. Hieruit blijkt dat vooral het algoritme met de *volledige regel* een duidelijk beter resultaat geeft dan alle andere methoden.

Naast deze basismethoden uit de literatuur beschouwen we ook Latente Dirichlet Allocatie (LDA), een recente en geavanceerde methode uit de literatuur [131]. LDA is een techniek uit de statistiek die gebruik maakt van latente variabelen om onderwerpen de modelleren. Dit gebeurt door een onderwerp voor te stellen als een verdeling over woorden. Een document is dan



Figuur 8.18: *Zuiverheid* en f-waarde in functie van het aantal clusters $|\,(\mathcal{E}_D)_1\,|$ voor LDA

een verdeling over onderwerpen. Voor elk van deze twee verdelingen wordt de Dirichlet-verdeling vooropgesteld. De parameters van deze verdeling kunnen worden geschat door een procedure die *Gibbs sampling* wordt genoemd. Aangezien LDA werkt op het niveau van woorden kunnen we geen PCA toepassen. Alle andere voorverwerkingstappen worden wel toegepast. Merk op dat in het LDA model $\Pr(\rho(d) = e)$ kan verschillen van nul voor meerdere e, d.i. een document kan worden gelinkt aan verschillende entiteiten. Om de *zuiverheid* en de f-waarde te kunnen berekenen kennen we aan elk document d het onderwerp met maximale waarschijnlijkheid toe. Uit Figuur 8.18 blijkt duidelijk dat LDA alle basismethoden overtreft. Het valt ook op dat de resultaten relatief stabiel zijn voor verschillende waarden van $|\,(\mathcal{E}_D)_1\,|$. Dit is, net als bij kernel *k-means* clusteren, te wijten aan het feit dat heel wat lege clusters worden gevormd indien het aantal vooraf bepaalde clusters te hoog is. Echter, LDA maakt een duidelijk betere invulling van de clusters dan kernel *k-means* clusteren. Dit maakt de LDA-methode een bijzonder uitdagend referentiepunt om onze aanpak aan te toetsen.

Wat betreft onze eigen methode, worden drie mogelijkheden voor selectie van documenten op basis van een patroon $(c_1, c_2) \in \mathcal{C}^2$ getest. Deze drie mogelijkheden voor selectie zijn gedefinieerd in Definities 8.8, 8.9 en 8.10. We zullen tijdens de rapportering naar onze methode refereren als het CR-clusteralgoritme (Concept Relationeel). We verwijzen naar één van de vermelde opties aan de hand van een superscript tussen haken. Bij $CR^{(1)}$ wordt gebruik gemaakt van de relationele selectie, bij $CR^{(2)}$ wordt gebruik gemaakt van de conceptuele selectie en bij $CR^{(3)}$ wordt gebruik gemaakt van de possibilistische selectie. We vergelijken onze aanpak met de maximale f-waarden die voor alternatieve methoden zijn vastgesteld overheen variaties in het aantal clusters en we rapporteren de bijhorende *zuiverheid*. Daarnaast rapporteren we ook het aantal clusters ($|\widehat{(\mathcal{E}_D)_1}|$) verkregen bij elke test. Tabel 8.2 toont de resultaten van deze vergelijking. Hieruit blijkt dat onze methode, voor wat betreft de *zuiverheid* van de clusters, een sterke verbetering biedt ten opzichte van zowel de basismethoden als de geavanceerde LDA methode.

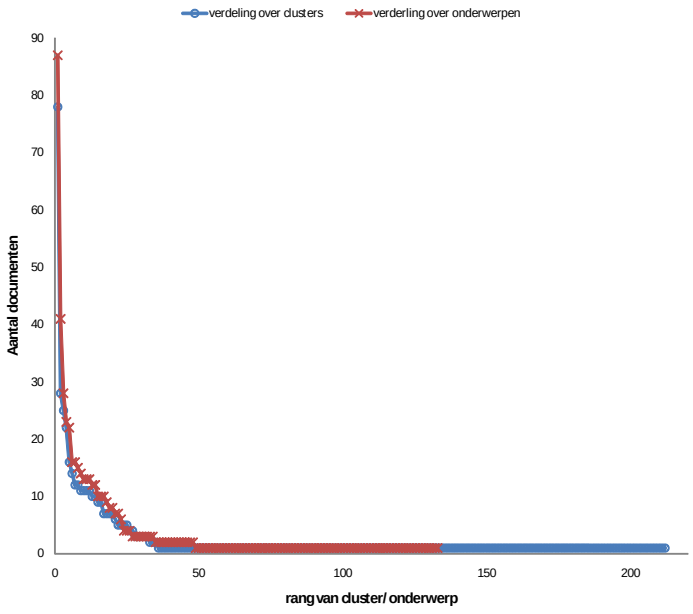
methode	zuiv	f	$\lvert \widehat{(\mathcal{E}_D)_1} \rvert$
$CR^{(1)}$	0.9508	0.6240	348
$CR^{(2)}$	0.9326	0.8036	219
$CR^{(3)}$	0.9399	0.8288	205
k-means	0.6904	0.5731	100
kernel *k-means*	0.6976	0.6481	49
hiërarchisch (enkel)	0.8925	0.4714	390
hiërarchisch (volledig)	0.8900	0.6927	240
LDA	0.8452	0.8107	92

Tabel 8.2: Resultaten van de nieuwe aanpak (CR) vergeleken met bestaande methoden

De *compleetheid* wordt in onze aanpak ondergeschikt aan de *zuiverheid*. Dit kan worden afgeleid uit de hoge *zuiverheid* in combinatie met een iets lagere f-waarde. De optie met de possibilistische selectie geeft duidelijk het beste resultaat op gebied van *compleetheid*, hoewel de conceptuele selectie bij benadering hetzelfde resultaat geeft. Ondanks het grotere belang dat wordt gegeven aan *zuiverheid*, wordt op gebied van *compleetheid* een duidelijke verbetering vastgesteld ten opzichte van bestaande methoden, zeker ten opzichte van de basismethoden. Ten opzichte van LDA is het verschil in f-waarde zeer klein. Laat ons de discussie daarom richten op een vergelijking van onze aanpak met LDA. Een opmerkelijke vaststelling is het verschil in het gegenereerde aantal clusters. Bij LDA ligt dit aantal duidelijk veel lager dan voor onze aanpak. Dit illustreert mooi het verschil tussen de twee methoden. Onze aanpak vertrekt van een controle van *zuiverheid*, d.i. documenten worden in clusters geplaatst en het aantal clusters wordt zo dicht mogelijk bij het geschatte aantal clusters gebracht. Echter, de *zuiverheid* wordt steeds gecontroleerd door een analyse van afhankelijkheden. Wanneer er wordt vastgesteld dat het verder clusteren de *zuiverheid* in het gedrang brengt, stopt onze aanpak. De strategie van LDA is duidelijk anders. Figuur 8.18 toont dat de f-waarde en de *zuiverheid* voor LDA stabiel zijn in functie van het opgegeven aantal clusters. De reden hiervoor is te vinden in het tweevoudig gebruik van de Dirichlet-verdeling, waardoor LDA impliciet een realistische schatting kan maken van het aantal onderwerpen. Echter, LDA houdt vast aan deze schatting en probeert dit aantal clusters zo optimaal mogelijk in te vullen, ondanks het eventuele verlies van *zuiverheid*. Een probleem hierbij is dat onzuivere clusters niet makkelijk op te kuisen zijn. Zuivere clusters met een lagere *compleetheid* kunnen echter wel makkelijk worden samengevoegd. Dit verschil in strategie speelt een belangrijke rol bij de verdere verwerking van clusters. Willen we bijvoorbeeld gaan naar een automatische samenvatting van teksten [119], dan zullen zuivere clusters steeds leiden tot coherente en leesbare samenvattingen. Onzuivere clusters zullen aanleiding geven tot incoherente en verwarrende samenvattingen omdat het onderwerp van een cluster niet duidelijk is afgelijnd. Het produceren van clusters met een hoge *zuiverheid* wordt dan ook beschouwd als een voordeel

van onze aanpak ten opzichte van LDA.

Om een antwoord te formuleren op de vraag hoe de *compleetheid* van onze aanpak kan worden verbeterd, beschouwen we Figuur 8.19. In deze figuur wordt de verdeling van documenten over entiteiten vergeleken met de verdeling van documenten over clusters. Hieruit blijkt dat de staart van de verdeling over clusters duidelijk veel langer is dan deze van de verdeling over entiteiten. Wanneer we dit verder analyseren, blijkt dat een groot deel van de singletonclusters door geen enkel patroon wordt geselecteerd. Er zijn dus heel wat documenten die geen veel voorkomende patronen bevatten, waardoor deze documenten buiten de patroonanalyse vallen. Deze documenten komen automatisch in een singletoncluster terecht. Een dergelijke vaststelling toont de grens aan van wat bereikt kan worden op basis van patronen. Dit sluit echter niet uit dat de basismethode bijvoorbeeld kan worden verfijnd met analyses van de concepten.



Figuur 8.19: Verdeling van de documenten over entiteiten en over clusters

In een volgend experiment wordt de nieuwe aanpak voor tekstclustering uitgebreider getest door de methode met de beste instellingen die in het vorige experiment gevonden is, toe te passen op willekeurig geselecteerde delen van de verzamelingen van documenten. Meer bepaald, voor steekproefgroottes 100, 200 en 300 worden telkens 50 steekproeven genomen. Elke steekproef wordt geclusterd en de gemiddelde resultaten worden berekend. Tabel 8.3 toont deze resultaten. Hierbij tonen de kolommen 'gem. $|(\mathcal{E}_D)_1|$' en 'gem. $|\widehat{(\mathcal{E}_D)_1}|$' respectievelijk het gemiddeld aantal onderwerpen en het gemiddeld aantal ge-

$\|D\|$	gem. $\|(\mathcal{E}_D)_1\|$	gem. $\|\widehat{(\mathcal{E}_D)_1}\|$	zuiv	f
100	44	75	0.9446	0.7066
200	71	118	0.9189	0.7176
300	90	158	0.8886	0.6944

Tabel 8.3: Gemiddelde resultaten voor verschillende steekproefgroottes met samenvoegingstap

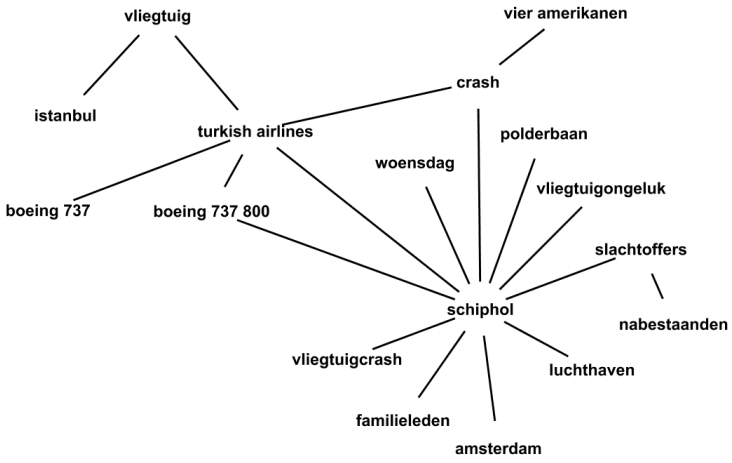
$\|D\|$	gem. $\|(\mathcal{E}_D)_1\|$	gem. $\|\widehat{(\mathcal{E}_D)_1}\|$	zuiv	f
100	46	74	0.9540	0.7262
200	70	125	0.9311	0.6955
300	90	162	0.9126	0.6946

Tabel 8.4: Gemiddelde resultaten voor verschillende steekproefgroottes zonder samenvoegingstap

genereerde clusters. Interessant om op te merken is dat het aantal clusters opnieuw heel wat hoger ligt dan het aantal onderwerpen. Dit ondanks het feit dat de schatting van het aantal clusters veel dichter bij het werkelijk aantal onderwerpen ligt. Opvallend is dat de *zuiverheid* en de f-waarde lager liggen dan op de volledige verzameling van documenten.

Merk op dat hoewel de f-waarde constant blijft (geen significante verschillen vastgesteld met een t-toets), dit niet het geval is voor de *zuiverheid*. Dit is verder onderzocht en het is gebleken dat het samenvoegen van clusters de oorzaak is van de duidelijk lagere *zuiverheid*. Tabel 8.4 toont de resultaten van hetzelfde experiment maar zonder de samenvoegingstap. De gemiddelde *zuiverheid* ligt duidelijk hoger dan wanneer samenvoeging gebruikt wordt na het clusteren. De f-waarde is ongeveer dezelfde, waaruit besloten kan worden dat de *compleetheid* lager moet zijn. Een bijzonder voordeel van onze aanpak is dat de patronen die worden gebruikt om een cluster te genereren, een representatieve skeletstructuur vormen die het samenvatten van de cluster vereenvoudigt. Figuur 8.20 toont de patronen die worden gebruikt om de cluster in verband met het vliegtuigongeluk in Schiphol te vormen. Uit deze patronen kan worden afgeleid dat het een vliegtuigongeluk op de luchthaven van Schiphol in Amsterdam op een woensdag betreft, waarbij een Boeing 737 van Turkish Airlines afkomstig uit Istanbul is neergestort nabij de Polderbaan. Door een terugkoppeling van een dergelijke structuur naar de i.Know technologie kunnen bovendien de relaties tussen concepten worden benoemd, zodat het automatisch genereren van een samenvatting opnieuw een stap dichterbij komt. Het samenvatten van objecten in het algemeen zal in Hoofdstuk 9 kort worden onderzocht.

Laat ons deze sectie afsluiten met enkele besluiten die uit de uitgevoerde experimenten kunnen worden getrokken. Het aantal entiteiten dat in een verzameling van documenten is vertegenwoordigd, kan via conceptanalyse op een vrij nauwkeurige manier worden geschat. Het is gebleken dat de nauwkeurig-



Figuur 8.20: Patronen als samenvatting van een cluster

heid ruimschoots volstaat in de context van onze aanpak voor tekstclustering. Het nieuwe algoritme voor het vinden van coreferente documenten genereert clusters met een hoge *zuiverheid*. De *compleetheid* van de verkregen clusters is minder hoog, maar in vergelijking met bestaande basistechnieken wordt een duidelijke en significante verbetering vastgesteld. Een belangrijke reden voor het niet-volledig zijn van de clusters is het feit dat niet alle documenten voldoende voorkomende patronen bezitten, waardoor deze documenten nooit kunnen worden geselecteerd. Dit is meteen de grens van de techniek gebaseerd op patronen. Belangrijk om te weten is dat het aantal verkregen clusters steeds duidelijk groter is dan het ingeschatte aantal. Het algoritme detecteert bijgevolg dat de gemaakte clusters te kampen hebben met een lagere *compleetheid*, maar het aanvaardt dit om de *zuiverheid* van de clusters niet te compromitteren. Door deze kennis kan het algoritme worden aangepast, zodat het teveel aan singletonclusters wordt weggewerkt. Hiervoor zijn uiteraard andere technieken nodig, die bijvoorbeeld gebaseerd kunnen zijn op een studie van concepten, eerder dan op een studie van de patronen (d.i. koppels van concepten). Dit onderzoek ligt echter buiten het bestek van dit werk.

8.8 Conclusie

In dit hoofdstuk is onderzocht hoe coreferentie kan worden behandeld wanneer de te vergelijken objecten tekstdocumenten zijn. Hierbij is vertrokken van een revolutionaire en gepatenteerde taaltechnologie, beschikbaar gesteld

door het bedrijf i.Know. Door gebruik te maken van de i.Know-technologie worden documenten niet gezien als een vector van woorden, maar eerder als een semantisch rijke structuur, bestaande uit concepten en relaties. In dit hoofdstuk is een nieuw documentmodel ontworpen, gebaseerd op multirelaties, dat kan worden ingevuld met deze technologie. Het is gebleken dat het construeren van een evaluator voor documenten in dit documentmodel wordt bemoeilijkt door het niet volledig gekend zijn van het meetproces. Op basis van een aantal veronderstellingen aangaande de evaluator voor documenten en het meetproces $\mathscr{M}$, kan het coreferentieprobleem worden geschreven in termen van een binaire relatie $R^{(\mathcal{D})}$ over een conceptruimte $\mathcal{C}$ en een selectieoperator voor tekstdocumenten. In deze notatie kunnen we een nieuwe oplossing voor tekstvergelijking formuleren door enerzijds een bepaling te maken van het aantal onderwerpen dat wordt vertegenwoordigd in een collectie van documenten en anderzijds een partitie van deze collectie af te leiden. Voor de bepaling van het aantal onderwerpen wordt verondersteld dat de verdeling van documenten over onderwerpen onderhevig is aan de Wet van Zipf. In een dergelijke context kan het aantal onderwerpen dat vertegenwoordigd is in een verzameling van documenten, accuraat worden bepaald door te steunen op het nieuwe documentmodel. Het schatten van het aantal onderwerpen is vervolgens gebruikt in een aanpak voor het vormen van een partitie van documenten, waarbij een verzameling documenten wordt voorgesteld als een multirelatie. Uit deze multirelatie kunnen verschillende sneden worden afgeleid, waarbij een snede steeds bestaat uit verschillende ongeconnecteerde en maximale deelcomponenten. Als er rekening wordt gehouden met afhankelijkheden binnen een component en afhankelijkheden tussen componenten, komen deze deelcomponenten ruwweg overeen met welbepaalde onderwerpen. De vernoemde afhankelijkheden kunnen in kaart worden gebracht door het gebruik van operatoren in het nieuwe documentmodel. Onze methode bestaat erin iteratief te zoeken naar een optimale snede, hetgeen een snede is waarvoor het aantal geproduceerde clusters zo dicht mogelijk bij het geschatte aantal clusters ligt. Experimenten tonen aan dat onze methode voor Nederlandstalige documenten bijzonder accurate clusters genereert in termen van *zuiverheid*. De *compleetheid* van clusters wordt ondergeschikt aan de *zuiverheid* van de clusters, maar de verkregen *compleetheid* van de clusters vertoont een duidelijke verbetering in vergelijking met bestaande methoden.

Hoofdstuk 9

Samenvoeging van objecten

9.1 Inleiding

In voorgaande hoofdstukken is het coreferentieprobleem bestudeerd voor verschillende gevallen. Hierbij komen we binnen de context van deze thesis steeds tot een partitie van het objectuniversum O. In dit hoofdstuk bestuderen we hoe objecten binnen een partitieklasse verder verwerkt kunnen worden. Meer specifiek zullen we onderzoeken hoe coreferente objecten kunnen worden samengevoegd. Dergelijke samenvoegingsprocessen zijn nuttig wanneer een groep van coreferente objecten moet worden gereduceerd tot één enkel object. Bijvoorbeeld in de context van een ETL proces moeten coreferente objecten uit de verschillende brondatabanken worden samengevoegd tot één consistent object dat in de *data warehouse* wordt geplaatst. We wijzen erop dat dit laatste hoofdstuk zeker geen volledige studie vormt van het samenvoegen van objecten. Eerder willen we in dit hoofdstuk nagaan in welke mate evaluatoren kunnen worden gebruikt in een context van samenvoeging. Dit hoofdstuk moet worden beschouwd als een exploratieve studie naar de mogelijkheden op het gebied van samenvoeging.

In Sectie 9.2 wordt een overzicht van relevante literatuur gegeven. In Sectie 9.3 wordt een algemene definitie voor samenvoegingsfuncties gegeven en worden relevante eigenschappen opgesomd. In Sectie 9.4 worden samenvoegingsfuncties voorgesteld voor atomaire objecten op basis van evaluatoren. Er wordt bestudeerd aan welke eigenschappen deze samenvoegingsfuncties voldoen. Daarna wordt in Sectie 9.5 bestudeerd hoe samenvoegingsfuncties voor complexe objecten kunnen worden geconstrueerd op basis van samenvoegingsfuncties voor atomaire objecten. Opnieuw wordt onderzocht welke eigenschappen voldaan zijn. Sectie 9.6 biedt een overzicht van de belangrijkste bevindingen uit dit hoofdstuk.

9.2 Overzicht van de literatuur

De samenvoeging van informatie behelst een ruim onderzoeksdomein dat reeds heel wat resultaten heeft opgeleverd. In het algemeen beschouwt men hierbij verschillende bronnen van informatie en zoekt men naar efficiënte voorstellingsvormen van de gezamenlijke kennis, waarbij eventuele inconsistenties tussen de verschillende bronnen moeten worden opgelost. Afhankelijk van hoe de kennis van bronnen wordt voorgesteld, leidt dit tot verschillende problemen met specifieke oplossingen. In deze sectie worden een aantal interessante resultaten uit de literatuur over samenvoeging van informatie besproken.

In een wiskundige context heeft samenvoeging van informatie geleid tot de ontwikkeling van een uitgebreid gamma van aggregatiefuncties zoals triangulaire normen en conormen [7], veralgemeende gemiddelden [132, 133] en uninormen [134]. Aggregatiefuncties worden steeds gedefinieerd voor een tralie $(L, \leq)$. De kennis die wordt voorgesteld geeft uitdrukking aan feiten. Beschouw als voorbeeld een beslissingsprobleem waarbij een panel van juryleden een beslissing moet nemen. Veronderstel hierbij een tralie $L = \{nee, neutraal, ja\}$ waarin elk jurylid zijn beslissing kan uitdrukken. Hoe luidt dan de beslissing van de gezamenlijke jury? Dit is een vraag die kan worden beantwoord door een gepaste aggregatiefunctie te gebruiken.

Naast aggregatiefuncties is heel wat onderzoek besteed aan de situatie waarbij een bron van informatie een propositionele kennisbank is, die wordt gemodelleerd als een eerste-orde theorie [135, 136, 137, 138, 139, 140]. Typisch bevatten dergelijke kennisbanken ook niet-feitelijke kennis zoals deductieregels en randvoorwaarden. In dergelijke situaties geeft de samenvoeging van informatie aanleiding tot kennis die in geen enkele van de bronnen afzonderlijk aanwezig is. Een relevante toepassing van deze problematiek zijn de zogenaamde heterogene databanken [141]. Dit zijn databanken waarbij gegevens worden verspreid over verschillende fysieke machines.

Voorbeeld 9.1
Laat $\mathscr{X}$ de verzameling van patiënten zijn en beschouw de verzameling van mogelijke symptomen $\{S_1, S_2, S_3\}$ en de verzameling van mogelijke diagnoses $\{D_1, D_2\}$. Laat voor alle $x \in \mathscr{X}$, $S_i(x)$ betekenen dat patiënt x symptoom S_i heeft en $D_i(x)$ dat voor patiënt x diagnose D_i wordt vastgesteld. Beschouw twee kennisbanken B_1 en B_2 zodat:

$$B_1 = \{S_1(a), S_3(b), \forall x \in \mathscr{X} : (S_1(x) \wedge S_2(x)) \Rightarrow D_2(x)\} \quad (9.1)$$

$$B_2 = \{S_2(a), \forall x \in \mathscr{X} : S_3(x) \Rightarrow D_1(x)\}. \quad (9.2)$$

Geen enkel van de kennisbanken doet rechtstreeks een uitspraak over diagnoses, maar het samenbrengen van kennis geeft aanleiding tot de diagnoses $D_2(a)$ en $D_1(b)$. Stel dat er bovendien een derde kennisbank

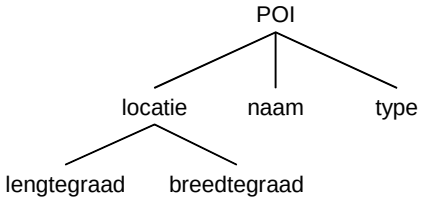
$$B_3 = \{\neg S_2(a)\} \quad (9.3)$$

voorhanden is. In dat geval is er een duidelijke tegenspraak aanwezig omtrent het vaststellen van symptoom S_2 bij patiënt a.

In een derde context waarin samenvoeging van informatie is onderzocht, wordt er verondersteld dat bronnen van informatie onzekerheid kunnen bevatten. Onder andere in [142] en [143] is onderzocht hoe samenvoeging kan plaatsvinden als informatiebronnen gebruik maken van possibiliteitsverdelingen om onzekerheid uit te drukken. Hierbij wordt vooral aandacht besteed aan de propagatie van de waargenomen onzekerheid.

9.3 Definities en eigenschappen

In deze sectie geven we eerst een zeer algemene definitie voor een samenvoegingsfunctie en introduceren we daarna één voor één een aantal eigenschappen die in bestaande toepassingen van informatiesamenvoeging hun nut hebben bewezen. In dit hoofdstuk worden geen experimenten gerapporteerd, voornamelijk omdat de correctheid van een resultaat in de context van samenvoeging niet eenvoudig objectief te beoordelen is. Meer bepaald is de constructie van grondwaarheid gevoelig voor interpretatie. Ter vervanging van experimenten zal met voorbeelden worden geïllustreerd hoe bepaalde samenvoegingsfuncties zich gedragen. Als lopend voorbeeld beschouwen we een website waarop geografische data worden verzameld volgens een *gemeenschapsgedreven* model. Dit wil zeggen dat gebruikers in staat zijn entiteiten aan te duiden op een kaart en deze objecten van metadata te voorzien. Enerzijds voor de eenvoud en anderzijds voor een overeenstemming met de realiteit, wordt een geografische locatie beschreven door een punt. Verder worden een naam en een type voorzien.



Figuur 9.1: Objectstructuur van POIs

Objecten die plaatsen beschrijven worden hier POIs (Points Of Interest) genoemd. In Voorbeelden 2.2 en 2.3 zijn de functies lab en λ voor POIs geïntroduceerd. Deze functies kunnen we schematisch voorstellen, zoals wordt getoond in Figuur 9.1. Het model waarbij gebruikers vrij POIs kunnen toevoegen aan een publieke verzameling zorgt ervoor dat heel wat coreferente POIs kunnen optreden. Op een verzameling van ongeveer 2500 POIs die locaties in Gent beschrijven, is de techniek voor objectvergelijking zoals beschreven in Hoofdstuk 7 toegepast. Als resultaat zijn een duizendtal koppels van POIs geïdentificeerd als zijnde coreferent. Door toepassing van de technieken in Hoofdstuk 5 komen we tot een partitie van de verzameling van POIs. Tabel 9.1 toont een aantal records, waarbij partitieklassen (d.z. groepen van coreferente objecten) afgescheiden staan.

naam	lengtegraad	breedtegraad	type
Huis van Alijn	3.723372	51.057411	Hist. Gebouw
Het huis van Alijn	3.723372	51.057411	Museum
Huis van Alijn	3.723504	51.057517	Algemeen
Het Huis van Alijn	3.722721	51.093481	Museum
Lakenhalle	3.725098	51.053552	Hist. Gebouw
Belfort	3.724837	51.053555	Algemeen
Belfort en Lakenhalle	3.724911	51.053653	Monument
SMAK	3.722726	51.038081	Algemeen
SMAK	3.722726	51.038081	Algemeen
Stedelijk Museum voor Actuele Kunst (SMAK)	3.723271	51.038129	Algemeen
SMAK - Stedelijk Museum voor Actuele Kunst	3.724141	51.038061	Museum

Tabel 9.1: Steekproef van POIs

We geven nu eerst de definitie van een samenvoegingsfunctie.

Definitie 9.1 (Samenvoegingsfunctie)
Gegeven een universum van objecten O, dan is een samenvoegingsfunctie over O gedefinieerd als:

$$\varpi_O : \mathcal{M}(O) \to O. \tag{9.4}$$

Een samenvoegingsfunctie beeldt een multiverzameling van objecten af op één enkel object. Definitie 9.1 is zeer algemeen en laat toe om samenvoegingsfuncties te definiëren die weinig zinvol zijn. Daarom wordt een opsomming gegeven van eigenschappen die een samenvoegingsfunctie zinvol maken. Een eerste eigenschap is idempotentie.

Eigenschap 9.1 (Idempotentie)
Een samenvoegingsfunctie ϖ_O is idempotent als:

$$\forall o \in O : \varpi_O(\{o, ..., o\}) = o. \tag{9.5}$$

Idempotentie wil zeggen dat een multiverzameling van objecten die onderling gelijk zijn, moet worden samengevoegd tot een object uit de multiverzameling. In bijna elke context van informatiesamenvoeging is idempotentie een natuurlijke eigenschap. In wat volgt wordt een samenvoegingsfunctie dan ook steeds als idempotent verondersteld. Een tweede interessante eigenschap, die is ontleend aan de definitie van aggregatiefuncties, is monotoniteit.

Eigenschap 9.2 (Monotoniteit)
Beschouw twee multiverzamelingen M_1 en M_2 die objecten uit O bevatten en beschouw een orderelatie $\leq$ over O. Beschouw een willekeurige één-op-één

afbeelding f tussen M_1 en M_2 zodat er geldt dat:

$$\forall (o_1, o_2) \in f : o_1 \leq o_2. \tag{9.6}$$

Een samenvoegingsfunctie ϖ_O is monotoon als er geldt dat:

$$\varpi_O(M_1) \leq \varpi_O(M_2). \tag{9.7}$$

Monotoniteit is in de context van coreferente objecten een natuurlijke en nuttige eigenschap. Echter, de orderelatie $\leq$ waarvan sprake, is niet altijd eenduidig te bepalen. In het geval van strings wordt de alfabetische ordening meestal als de natuurlijke ordening beschouwd, maar in de context van samenvoeging zijn de partiële orderelaties $\sqsubset$ en $\widehat{\sqsubset}$ (Hoofdstuk 6) eveneens nuttig. Ook voor complexe objecten is het niet meteen duidelijk hoe een dergelijke orderelatie moet worden opgebouwd. We zullen monotoniteit in wat volgt niet nader onderzoeken, maar we vermelden deze eigenschap omdat ze nuttig kan zijn voor samenvoeging van numerieke objecten. Een derde eigenschap is bewaring.

Eigenschap 9.3 (Bewaring)
Een samenvoegingsfunctie ϖ_O is bewarend als:

$$\forall M \in \mathcal{M}(O) : \varpi_O(M) \in M. \tag{9.8}$$

Een bewarende samenvoegingsfunctie beeldt een multiverzameling af op een element van de multiverzameling en is dus altijd idempotent. Bewaring is in de context van coreferente objecten een bijzonder interessante eigenschap om verschillende redenen. Ten eerste laat een bewarende samenvoegingsfunctie traceerbaarheid toe. Dit betekent dat voor een samengevoegd object altijd kan worden afgeleid welke bron dit object bevat. In vele praktische situaties is dit een nuttige eigenschap. Denken we bijvoorbeeld aan de context van *data warehousing*, dan kan traceerbaarheid een middel bieden om te achterhalen welke van de aanwezige bronnen typisch worden geselecteerd om de *data warehouse* te voeden. Ten tweede, in het geval van strings is bewaring veelal een middel om te verzekeren dat het samengevoegde object een zinvol object is, dat duidelijk verwijst naar de entiteit onder beschouwing. In het geval van de eigenschap 'naam', is bewaring bijvoorbeeld een eigenschap die bijzonder intuïtief aanvoelt. Ten derde, voor complexe objecten kan een willekeurig mengsel van deelobjecten onzinvol zijn. Wanneer men bijvoorbeeld twee adressen wil samenvoegen zal men niet de straatnaam van het eerste adres en het huisnummer van het tweede adres nemen, aangezien er geen verificatie is dat het resulterende adres bestaat.

Een dieper inzicht in het nut van bewaring kan worden verkregen wanneer de creatie van objecten wordt beschouwd als een meetproces (Figuur 2.2). In Hoofdstuk 2 is aangehaald hoe moeilijk-meetbare eigenschappen kunnen worden behandeld tijdens het zoeken naar coreferente objecten. In die uiteenzetting is uitgegaan van een possibilistische omgeving. In realiteit is het echter niet ondenkbaar en zelfs zeer waarschijnlijk dat een dergelijke possibilistische

omgeving niet wordt gebruikt. Ten gevolge hiervan worden moeilijk-meetbare eigenschappen beschreven door objecten alsof het perfect meetbare eigenschappen zijn. Wanneer we dergelijke objecten gaan samenvoegen, bestaat er een grote zekerheid dat geen enkel van de objecten een correcte beschrijving geeft van de entiteit in kwestie. Bewaring is in die context geen eigenschap die de kwaliteit van het resultaat verhoogt. In het geval van POIs zijn 'lengtegraad' en 'breedtegraad' moeilijk-meetbare eigenschappen. De eigenschap 'naam' is een meetbare eigenschap. In die zin is bewaring vooral nuttig bij het samenvoegen van de eigenschap 'naam'.

Een vierde eigenschap, ontleend aan de literatuur rond propositionele kennisbanken, is de zogenaamde meerderheidsregel [138, 140], die we hier als volgt definiëren:

Eigenschap 9.4 (Meerderheidsregel)
Een samenvoegingsfunctie ϖ_O voldoet aan de meerderheidsregel als:

$$\forall M \in \mathcal{M}(O) : \exists o \in O : \omega_M(o) > \frac{|M|}{2} \Rightarrow (\varpi_O(M) = o)\,. \tag{9.9}$$

De dominantie van een absolute meerderheid kan een nuttige eigenschap zijn aangezien het uitdrukking geeft aan een natuurlijk keuzecriterium. Echter, de meerderheidsregel leidt tot een typische foutpropagatie wanneer er sterke afhankelijkheden bestaan tussen de verschillende bronnen. Om dit verduidelijken, beschouwen we het voorbeeld van verwijzingen in wetenschappelijke artikels. Wanneer coreferente verwijzingen worden samengevoegd, is het belangrijk te weten dat een groot deel van verwijzingen onrechtstreeks verlopen. De verwijzing van artikel x naar artikel z wordt bijvoorbeeld overgenomen van artikel y. In een dergelijke ketting van verwijzingen worden fouten typisch gemakkelijk gepropageerd waardoor een meerderheid aan foute verwijzingen kan ontstaan. De meerderheidsregel zal dan leiden tot een foute verwijzing als resultaat van de samenvoeging. Daarom kan het nuttig zijn dat een samenvoegingsfunctie niet afhankelijk is van de multipliciteit van de objecten die worden samengevoegd. Om deze eigenschap te definiëren, voeren we eerst het principe van onderdrukking van een multiverzameling in.

Definitie 9.2 (k-onderdrukking)
Laat M een multiverzameling in U zijn, dan is de k-onderdrukking van M een multiverzameling $\langle M \rangle_k$ zodat er geldt dat:

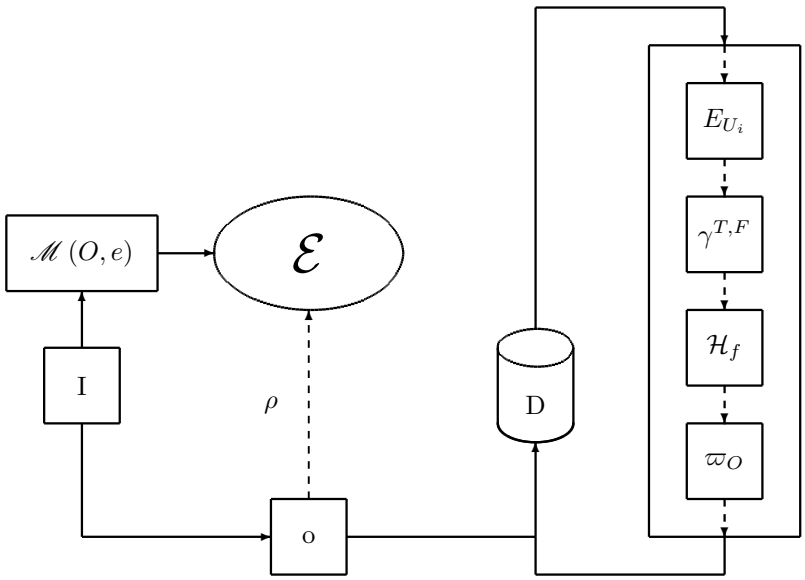
$$\omega_{\langle M \rangle_k}(u) = \begin{cases} 1 & \textbf{\textit{als}} \quad 0 < \omega_M(u) < k \\ \omega_M(u) & \textbf{\textit{anders}}. \end{cases} \tag{9.10}$$

Dit leidt tot de volgende definitie.

Eigenschap 9.5 (Meerderheidsonafhankelijkheid)
Een samenvoegingsfunctie ϖ_O is k-meerderheidsonafhankelijk als:

$$\forall M \in \mathcal{M}(O) : \varpi_O(M) = \varpi_O(\langle M \rangle_k). \tag{9.11}$$

Laat ons deze sectie besluiten met de situering van samenvoeging ten opzichte van coreferentiebepaling. Het mechanisme van samenvoeging kan namelijk worden gezien als een proces dat volgt op de bepaling van coreferentie. Dit wordt schematisch weergegeven in Figuur 9.2. Hierbij beschouwen we opnieuw een databank met objecten, waarbij we stapsgewijs een evaluator E_O voor complexe objecten construeren zoals beschreven in Hoofdstuk 7. Na het verkrijgen van een partitie door toepassing van $\mathcal{H}_f$, kunnen we een samenvoegingsfunctie ϖ_O toepassen op elk van de partitieklassen. Dit levert een opgekuiste databank zonder coreferente objecten.



Figuur 9.2: Coreferentiebepaling en samenvoeging

9.4 Samenvoeging van atomaire objecten

9.4.1 Het algemene geval

In deze sectie worden samenvoegingsfuncties ϖ_U bestudeerd, waarbij U een atomair universum voorstelt. Gelet op de context van coreferentie kunnen we veronderstellen dat een evaluator E_U beschikbaar is. Onze aandacht gaat daarom uit naar samenvoegingsfuncties die gebruik maken van een evaluator. Laat M een multiverzameling van coreferente (atomaire) objecten zijn. Voor een willekeurig object $u \in M$ kan u worden vergeleken met elk object in M, zodat $|M|$ possibilistische waarheidswaarden worden verkregen. Omwille van de reflexiviteit van E_U zal de possibilistische waarheidswaarde $(1, 0)$ minstens

$\omega_M(u)$ keer vookomen. De collectie van possibilistische waarheidswaarden die op deze manier voor een element $u \in M$ wordt verkregen, kan worden omgezet naar een possibiliteitsverdeling over $\mathbb{N}$. De vaagverzameling die een dergelijke possibiliteitsverdeling modelleert wordt in de literatuur een vaag natuurlijk getal genoemd [7].

Definitie 9.3
Een vaag natuurlijk getal is een genormaliseerde vaagverzameling over $\mathbb{N}$. *Een willekeurig vaag natuurlijk getal wordt genoteerd als* $\widetilde{n}$ *en de verzameling van alle vage natuurlijke getallen wordt genoteerd als* $\mathcal{F}(\mathbb{N})$.

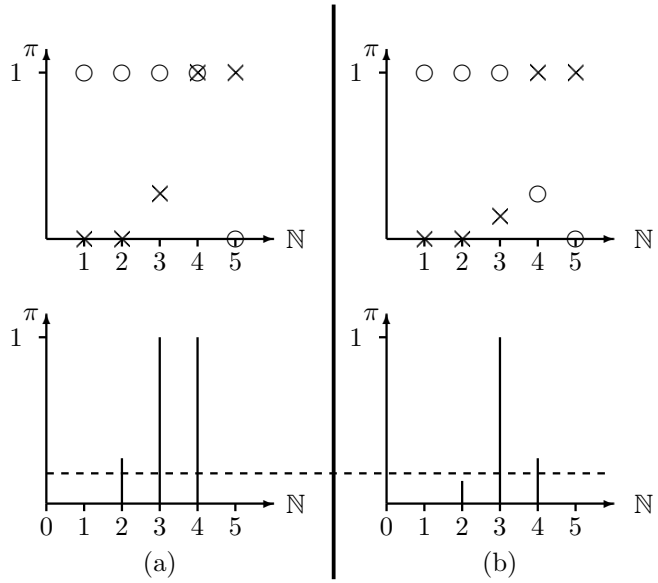
Voor de omzetting van een collectie possibilistische waarheidswaarden naar een vaag natuurlijk getal, steunen we op de methode van Hallez [144], die als volgt kan worden samengevat. Beschouw een Boolese propositie p, een verzameling $A = \{\mathcal{A}_1, ..., \mathcal{A}_n\}$ van actoren en laat P_π de overeenkomstige verzameling van possibilistische waarheidswaarden zijn, d.i. $P_\pi = \{\widetilde{p}_1, ..., \widetilde{p}_n\}$. Het aantal actoren in A dat $p = T$ postuleert volgens P_π is gegeven door de possibiliteitsverdeling $\pi_\mathbb{N}$ waarvoor er geldt dat:

$$\pi_\mathbb{N}(k) = \min \left(\begin{array}{ll} \sup & \left\{ \alpha \middle| \alpha \in [0,1] \wedge \left| \{\mathcal{A}_i | \mathcal{A}_i \in A \wedge \mu_{\widetilde{p}_i}(T) \geq \alpha\} \right| \geq k \right\}, \\ \sup & \left\{ \alpha \middle| \alpha \in [0,1] \wedge \left| \{\mathcal{A}_i | \mathcal{A}_i \in A \wedge \mu_{\widetilde{p}_i}(F) < \alpha\} \right| \geq k \right\} \end{array} \right) \tag{9.12}$$

waarbij $\widetilde{p}_i$ de onzekerheid over de waarheidswaarde van p volgens $\mathcal{A}_i$ voorstelt. Voor elke $k \in \mathbb{N}$ is $\pi_\mathbb{N}(k)$ de mogelijkheid dat exact k actoren $p = T$ postuleren. Deze mogelijkheid is gelijk aan het minimum van (1) de mogelijkheid dat minstens k actoren $p = T$ postuleren en (2) de mogelijkheid dat ten meeste $|A| - k$ actoren $p = F$ postuleren. In wat volgt zullen we deze interessante methode gebruiken voor de constructie van samenvoegingsfuncties. We geven daarom eerst enkele bijkomende eigenschappen van deze methode. De verdeling $\pi_\mathbb{N}$ kan namelijk bijzonder efficiënt worden berekend als volgt:

$$\pi_\mathbb{N}(k) = \left\{ \begin{array}{lll} \mu_{\widetilde{p}_{(k)^T}}(F) & \textbf{als} & k = 0 \\ \mu_{\widetilde{p}_{(k)^T}}(T) & \textbf{als} & k = |M| \\ \min\left(\mu_{\widetilde{p}_{(k)^T}}(T), \mu_{\widetilde{p}_{(k+1)^T}}(F)\right) & \textbf{anders}. & \end{array} \right. \tag{9.13}$$

waarbij $\cdot_{(i)^T}$ de ordening van possibilistische waarheidswaarden voorstelt zoals beschreven in Hoofdstuk 3. Wanneer de verzameling P_π gesorteerd is, heeft de constructie van $\pi_\mathbb{N}$ bijgevolg een lineaire complexiteit. Het kan nu ook gemakkelijk worden ingezien dat de vaagverzameling die $\pi_\mathbb{N}$ modelleert, steeds convex is (d.i. een convexe lidmaatschapsfunctie heeft) en dat $\pi_\mathbb{N}$ steeds een genormaliseerde verdeling is. Figuur 9.3 toont twee voorbeelden van een verzameling P_π. Een possibilistische waarheidswaarde wordt voorgesteld door een koppel van $\circ$ (mogelijkheid voor T) en $\times$ (mogelijkheid voor F). De afgeleide verdelingen $\pi_\mathbb{N}$ worden onderaan getoond.



Figuur 9.3: Afgeleide possibiliteitsverdelingen

Door toepassing van de methode van Hallez [144] kunnen we voor een element $u \in M$, uitdrukking geven aan de onzekerheid over het aantal objecten in M waarmee u coreferent is volgens de evaluator E_U. We kunnen nu op zoek gaan naar het object $u \in M$ dat volgens E_U coreferent is met een maximaal aantal objecten in M. Op basis van de kennis van E_U is dit object immers de meest geschikte voorstelling van de entiteit waarnaar wordt verwezen. De keuze tussen objecten wordt op deze manier vertaald naar een keuze tussen vage natuurlijke getallen. Een veel gebruikte methode om een dergelijke keuze te maken is de verscherping van de vage natuurlijke getallen, waarbij elk vaag natuurlijk getal wordt vertaald naar een (reëel) getal. Veelal wordt hiervoor de abscis van het zwaartepunt van de lidmaatschapsfunctie berekend [7]:

$$\frac{\sum_{k=0}^{|M|} k\ \pi_{\mathbb{N}}(k)}{\sum_{k=0}^{|M|} \pi_{\mathbb{N}}(k)}. \tag{9.14}$$

Het vergelijken van vage natuurlijke getallen gebeurt dan door vergelijking van de zwaartepunten van de lidmaatschapsfuncties. In dit hoofdstuk wensen we een methode te gebruiken die meer aanleunt bij de possibilistische denkwereld. We beschouwen twee orderelaties voor vage natuurlijke getallen:

Definitie 9.4 (sup**-orderelatie voor vage natuurlijke getallen)**
De orderelatie $\prec_{\text{sup}}$ voor vage natuurlijke getallen is gedefinieerd als:

$$\forall(\widetilde{n}, \widetilde{m}) \in \mathcal{F}(\mathbb{N})^2 : \widetilde{n} \prec_{sup} \widetilde{m} \Leftrightarrow \sup \widetilde{n}_\alpha < \sup \widetilde{m}_\alpha \tag{9.15}$$

waarbij $\widetilde{n}_\alpha$ de α-snede van $\widetilde{n}$ voorstelt en waarbij α wordt gekozen als volgt:

$$\alpha = \sup\{x | \sup \widetilde{n}_x \neq \sup \widetilde{m}_x\}. \tag{9.16}$$

Definitie 9.5 (inf**-orderelatie voor vage natuurlijke getallen)**
De orderelatie $\prec_{\text{inf}}$ voor vage natuurlijke getallen is gedefinieerd als:

$$\forall(\widetilde{n}, \widetilde{m}) \in \mathcal{F}(\mathbb{N})^2 : \widetilde{n} \prec_{\text{inf}} \widetilde{m} \Leftrightarrow \inf \widetilde{n}_\alpha < \inf \widetilde{m}_\alpha \tag{9.17}$$

waarbij $\widetilde{n}_\alpha$ de α-snede van $\widetilde{n}$ voorstelt en waarbij α wordt gekozen als volgt:

$$\alpha = \sup\{x | \inf \widetilde{n}_x \neq \inf \widetilde{m}_x\}. \tag{9.18}$$

De sup-orderelatie voor vage natuurlijke getallen zoekt naar de grootste α waarvoor de α-snedes een verschillend supremum hebben. Het grootste vage natuurlijk getal is dit met het grootste supremum van de bewuste α-snede. Dit betekent dat de lidmaatschapsgraden van de getallen die kleiner zijn dan het supremum van de 1-snede geen invloed hebben op de sup-ordening van vage natuurlijke getallen. Analoog zoekt de inf-orderelatie voor vage natuurlijke getallen naar de grootste α waarvoor de α-snedes een verschillend infimum hebben. Het grootste vage natuurlijk getal is dit met het grootste infimum van de bewuste α-snede. Dit betekent dat de lidmaatschapsgraden van de getallen die groter zijn dan het infimum van de 1-snede geen invloed hebben op de inf-ordening van vage natuurlijke getallen. Er volgt meteen dat $\prec_{\text{sup}}$ en $\prec_{\text{inf}}$ partiële orderelaties zijn. Beschouw als voorbeeld de vage natuurlijke getallen in Figuur 9.3. Onder de orderelatie $\prec_{\text{sup}}$ is het linkse vage natuurlijk getal groter, aangezien de 1-snede van het linkse vage natuurlijk getal een groter supremum (4) heeft dan het rechtse vage natuurlijk getal (3). Echter, onder de orderelatie $\prec_{\text{inf}}$ is het rechtse vage natuurlijk getal groter, aangezien de 0.2-snede (stippellijn) van het linkse vage natuurlijk getal het kleinste infimum heeft. We kunnen de ingevoerde orderelaties gebruiken om samenvoegingsfuncties te construeren.

Definitie 9.6 (Evaluator-gebaseerde samenvoeging)
Laat U een atomair universum zijn en laat E_U een evaluator over U zijn. Een samenvoegingsfunctie van orde k, gebaseerd op de evaluator E_U, is een samenvoegingsfunctie ϖ_U^k waarvoor er geldt dat:

$$\varpi_U^k(M) = \arg\max_{u \in M} \pi_{\mathbb{N}}^u \tag{9.19}$$

waarbij $\pi_{\mathbb{N}}^u$ een vaag natuurlijk getal is, verkregen uit de multiverzameling van possibilistische waarheidswaarden $P_{\pi,u}$ zodat:

$$\forall u' \in M : \omega_{P_{\pi,u}}(E_U(u, u')) = \omega_{\langle M \rangle_k}(u'). \tag{9.20}$$

Aangezien ϖ_U^k een element uit M onderdrukt als de multipliciteit kleiner is dan k, is de samenvoegingsfunctie ϖ_U^k k-meerderheidsonafhankelijk. Samenvoegingsfuncties ϖ_U^k zoals gedefinieerd in Definitie 9.6 zijn bij definitie bewarend

en dus ook idempotent. In wat volgt zullen we voor de eenvoud veronderstellen dat:

$$\varpi_U(M) = \varpi_U^1(M). \tag{9.21}$$

Het kan met tegenvoorbeelden worden aangetoond dat ϖ_U niet altijd aan de meerderheidsregel voldoet. Het blijkt zelfs dat de condities waaronder de meerderheidsregel wel geldt, vrij streng zijn. Dit valt af te leiden uit de volgende stelling.

Stelling 9.1
Een samenvoegingsfunctie ϖ_U, gebaseerd op een sterk reflexieve en transitieve evaluator E_U, voldoet aan de meerderheidsregel als de vage natuurlijke getallen worden geordend met de inf*-orderelatie.*

Bewijs. Stel dat a het element van M is dat de meerderheid heeft. Dit betekent dat:

$$\omega_M(a) > \left\lfloor \frac{|M|}{2} \right\rfloor. \tag{9.22}$$

Er geldt dat:

$$\forall u \in M : \left| \{\widetilde{p} | \widetilde{p} \in P_{\pi,u} \wedge \mu_{\widetilde{p}}(F) \neq 1\} \right| = \min\{k | k \in \mathbb{N} \wedge \pi_{\mathbb{N}}^u(k) = 1\} \tag{9.23}$$

wat betekent dat het infimum van de 1-snede van $\pi_{\mathbb{N}}$ gelijk is aan het aantal possibilistische waarheidswaarden met possibiliteit voor vals strikt kleiner dan 1. Als er geldt dat:

$$\forall u \in M \backslash \{a\} : \mu_{E_U(u,a)}(T) < 1 \tag{9.24}$$

dan heeft $\pi_{\mathbb{N}}^a$ bijgevolg het strikt grootste infimum voor zijn 1-snede onder alle elementen in M. Zoniet, dan bestaat er een multiverzameling $C \subset M$ zodat:

$$\forall u \in C : \mu_{E_U(u,a)}(T) = 1. \tag{9.25}$$

In dit geval geldt er wegens het eerste geval dat:

$$\forall u \notin C : \pi_{\mathbb{N}}^u \prec_{\text{inf}} \pi_{\mathbb{N}}^a \tag{9.26}$$

Voor elementen in C geldt er:

$$\forall u \in C : \pi_{\mathbb{N}}^u(|C|) = 1. \tag{9.27}$$

Het volstaat dan om aan te tonen dat voor elke $u \in C$, $\pi_{\mathbb{N}}^a$ puntsgewijs kleiner is dan $\pi_{\mathbb{N}}^u$, voor de indexverzameling $\{1, ..., |C|\}$. Enerzijds geldt er voor a:

$$\forall k \in \{1, ..., \lfloor |M|/2 \rfloor\} : \pi_{\mathbb{N}}^a(k) = 0. \tag{9.28}$$

Anderzijds geldt er voor een willekeurige $b \in C$ verschillend van a dat:

$$\omega_M(b) < \left\lfloor \frac{|M|}{2} \right\rfloor. \tag{9.29}$$

Aangezien E_U sterk reflexief is betekent dit dat:

$$\forall b \in C \backslash \{a\} : \exists k \in \{1, ..., \lfloor |M|/2 \rfloor\} : \pi^b_{\mathbb{N}}(k) > 0. \tag{9.30}$$

Dit betekent dat:

$$\forall b \in C \backslash \{a\} : \forall k \in \{1, ..., \lfloor |M|/2 \rfloor\} : \pi^a_{\mathbb{N}}(k) \leq \pi^b_{\mathbb{N}}(k). \tag{9.31}$$

Bovendien bestaat er minstens één index $k \in \{1, ..., \lfloor |M|/2 \rfloor\}$ waarvoor $\pi^a_{\mathbb{N}}(k)$ strikt kleiner is dan $\pi^b_{\mathbb{N}}(k)$. Met betrekking tot de indexverzameling $\{\lfloor |M|/2 \rfloor + 1, ..., |C|\}$ merken we op dat er voor elke b en c in C, verschillend van a, door de transitiviteit van E_U, geldt dat:

$$(E_U(b, c) < E_U(a, c)) \Rightarrow (E_U(a, b) = E_U(a, c)) \,. \tag{9.32}$$

Aangezien er geldt dat:

$$\forall b \in C \backslash \{a\} : \omega_M(b) < \left\lfloor \frac{|M|}{2} \right\rfloor \tag{9.33}$$

volgt hieruit dat:

$$\forall k \in \left\{ \left\lfloor \frac{|M|}{2} \right\rfloor + 1, ..., |C| \right\} : \pi^a_{\mathbb{N}}(k) \leq \pi^b_{\mathbb{N}}(k). \tag{9.34}$$

□

Voorbeeld 9.2
Laat ons het samenvoegen van atomaire objecten, gebaseerd op een evaluator, illustreren voor de eigenschap 'naam' voor de POIs uit Tabel 9.1, waarbij we een twee-niveau evaluator $E^*_{\mathcal{S}}$ beschouwen met parameters $(1, 0, 0.05)$ (Hoofdstuk 6). Tabel 9.2 toont de resultaten voor samenvoeging van de namen.

$\prec_{\text{inf}}$	$\prec_{\text{sup}}$
Huis van Alijn	Huis van Alijn
Belfort en Lakenhalle	Belfort en Lakenhalle
SMAK	SMAK

Tabel 9.2: Samenvoeging van de namen met $\varpi_{\mathcal{S}}$ gebaseerd op $E^*_{\mathcal{S}}$

9.4.2 Bijzondere gevallen

Enkele interessante samenvoegingsfuncties worden verkregen wanneer bijzondere evaluatoren worden beschouwd. Als de evaluator tweewaardig is, dan geldt de meerderheidsregel, los van de gebruikte orderelatie voor vage natuurlijke getallen. Als de evaluator driewaardig is, dan geldt de meerderheidsregel als de vage natuurlijke getallen vergeleken worden met de inf-orderelatie, behalve voor $\perp$. Dit wil zeggen dat wanneer in een multiverzameling $M \in \mathcal{M}(U)$ objecten

zitten die een niet-meetbare eigenschap beschrijven, deze objecten nooit het resultaat van $\varpi_U(M)$ kunnen zijn, op voorwaarde dat er minstens één object in M aanwezig is die een meetbare eigenschap beschrijft. Anders gezegd, het resultaat van $\varpi_U(M)$ is $\perp$ als en alleen als alle waarden uit M gelijk zijn aan $\perp$.

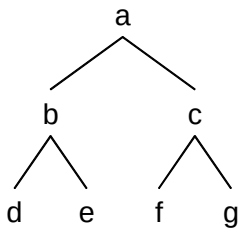
Een belangrijke eigenschap van bewarende samenvoegingsfuncties is de voorwaarden waaronder ze onbeslist zijn, d.i. geen keuze kunnen maken uit de gegeven multiverzameling van objecten M. Wanneer de samenvoegingsfunctie gebaseerd is op een evaluator E_U, dan kan onbeslistheid twee oorzaken hebben:

1. de hoogst geordende vage natuurlijke getallen zijn gelijk aan elkaar

2. de hoogst geordende vage natuurlijke getallen zijn verschillend, maar kunnen niet onderscheiden worden door de gebruikte orderelatie

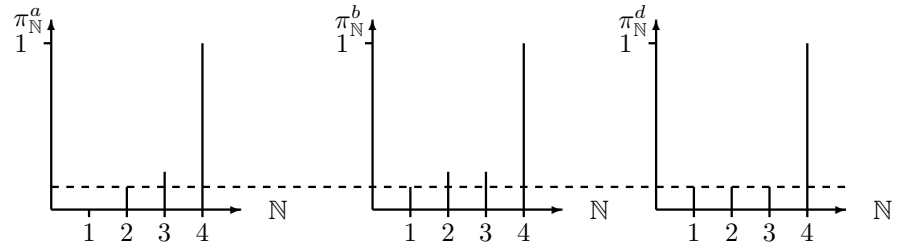
Het eerste geval treedt bijvoorbeeld op als M twee elementen met eenzelfde multipliciteit bevat en geen andere elementen dan deze twee. In dit geval kan enkel een willekeurige keuze worden gemaakt. In het tweede geval kunnen de hoogst geordende vage getallen worden vergeleken met een andere orderelatie, die de vage natuurlijke getallen wel kan onderscheiden. In het kader van onbeslistheid zijn semantische evaluatoren (Hoofdstuk 2) interessante gevallen. Meer bepaald kan worden vastgesteld dat een semantische evaluator aanleiding geeft tot de meest specifieke samenvoeging.

Voorbeeld 9.3
Beschouw een universum $U = \{a, b, c, d, e, f, g\}$ waarvoor een binaire relatie R beschikbaar is, die wordt afgebeeld in Figuur 9.4.



Figuur 9.4: Voorbeeld van een boomstructuur

Stel dat voor het universum U een kardinaliteitsgebaseerde *semantische evaluator $E_{U,R}$ met $g = \min$ (zie (2.55)) wordt gebruikt en beschouw de daarop gebaseerde samenvoegingsfunctie ϖ_U. Stel dat een multiverzameling $M =$*



Figuur 9.5: Vage natuurlijke getallen met semantische evaluatie

$\{a, a, b, d\}$ *gegeven is. De semantische evaluaties worden berekend als volgt:*

$$E_{U,R}(a,b) = \left(1, \min\left(\frac{6}{12}, \frac{3}{12}\right)\right) = (1, 0.25) \tag{9.35}$$

$$E_{U,R}(a,d) = \left(1, \min\left(\frac{6}{12}, \frac{2}{12}\right)\right) = (1, 0.17) \tag{9.36}$$

$$E_{U,R}(b,d) = \left(1, \min\left(\frac{3}{12}, \frac{2}{12}\right)\right) = (1, 0.17). \tag{9.37}$$

Dit geeft aanleiding tot de drie vage natuurlijke getallen getoond in Figuur 9.5. De 0.17*-snede (stippellijn) toont hierbij dat gebruik van de* inf*-orderelatie leidt tot* $\varpi_U(M) = d$*. De reden dat deze keuze logisch is, komt voort uit het feit dat* d *het meest specifieke object is, d.i.* d *staat het minst in relatie met andere objecten. Echter, dit neemt niet weg dat* $E_{U,R}$ *nog steeds onbeslist zal zijn voor* $M = \{a, d\}$*. Bovendien hangt het gedrag van* ϖ_U *af van de functie* g *die wordt gekozen om* $E_{U,R}$ *te berekenen.*

Indien een relatie R voorhanden is, kan het samenvoegen dus een stuk efficiënter en beter gebeuren door rechtstreeks op de relatie te steunen. Dit geeft aanleiding tot de definitie van een semantische samenvoegingsfunctie.

Definitie 9.7 (Semantische samenvoeging)
Laat U *een atomair universum zijn en laat* R *een binaire relatie over* U *zijn. Een semantische samenvoegingsfunctie gebaseerd op de relatie* R *is een samenvoegingsfunctie* $\varpi_{U,R}$ *zodat:*

$$\varpi_{U,R}(M) = \arg\min_{u \in M} |\operatorname{sel}_u(R)|. \tag{9.38}$$

Semantische samenvoeging is bij definitie altijd bewarend. Semantische samenvoeging is steeds k-meerderheidsonafhankelijk met:

$$k = \max_{u \in U}(\omega_M(u)). \tag{9.39}$$

Een semantische samenvoegingsfunctie is onbeslist als er minstens twee elementen uit M bestaan die $|\operatorname{sel}_u(R)|$ minimaliseren. Dit betekent dat indien M

bestaat uit twee elementen met eenzelfde multipliciteit en geen andere elementen, $\varpi_{U,R}(M)$ niet noodzakelijk onbeslist is, in tegenstelling tot samenvoeging gebaseerd op een evaluator.

Tot slot van deze sectie geven we een korte toelichting bij het samenvoegen van collecties. In de opsomming van eigenschappen is aangehaald dat bewaring een nuttige eigenschap kan zijn, bijvoorbeeld omwille van traceerbaarheid. Het kan echter worden ingezien dat deze eigenschap een vrij zware beperking oplegt aan samenvoegingsfuncties. In Hoofdstuk 8 is bijvoorbeeld het geval van teksten besproken. Als we beschikken over een collectie van teksten over hetzelfde onderwerp, dan kan samenvoeging van deze teksten worden gezien als tekstsamenvatting van meerdere documenten [119]. Bewaring leidt in deze context echter tot de beperking dat een samenvatting van meerdere teksten gelijk moet zijn aan één van de teksten. Gelet op het feit dat we in Hoofdstuk 8 een tekst (document) hebben voorgesteld als een multirelatie van concepten (d.i. een collectie van conceptkoppels), zou het voorzien van een soepelere eigenschap voor het geval van collecties bijzonder nuttig zijn. Daarom introduceren we in het geval van collecties de eigenschap $\subseteq$-bewaring.

Eigenschap 9.6 ($\subseteq$-bewaring)
Gegeven een universum U, dan wordt een samenvoegingsfunctie $\varpi_{\mathcal{M}(U)}$ $\subseteq$-bewarend genoemd als:

$$\forall M \in \mathcal{M}(\mathcal{M}(U)) : \forall o \in M : \varpi_{\mathcal{M}(U)}(M) \subseteq o \tag{9.40}$$

hetgeen equivalent is aan:

$$\forall M \in \mathcal{M}(\mathcal{M}(U)) : \varpi_{\mathcal{M}(U)}(M) \subseteq \bigcap_{o \in M} (o). \tag{9.41}$$

De eigenschap $\subseteq$-bewaring geeft een soepelere voorwaarde voor samenvoegingsfuncties. Het kan worden ingezien dat $\subseteq$-bewaring slechts een voorbeeld van een dergelijke eigenschap is en dat vele andere eigenschappen gedefinieerd kunnen worden. Een verdere studie hiervan valt buiten het bestek van deze thesis, maar is zeker interessant om verder te onderzoeken, ondermeer omwille van de formele onderbouw voor samenvatting van teksten die kan worden verkregen.

9.5 Samenvoeging van complexe objecten

In deze sectie worden samenvoegingsfuncties voor complexe objecten onderzocht. Een eerste manier om complexe objecten samen te voegen is het volgen van een zelfde redenering als bij het samenvoegen van atomaire objecten. Dit heeft echter het belangrijke nadeel dat een evaluator voor complexe objecten niet aan elke eigenschap eenzelfde conditionele necessiteit toekent. Het kan zelfs zijn dat bepaalde eigenschappen geen rol spelen in het bepalen van het resultaat van E_O. Daar waar tijdens het zoeken naar coreferente objecten verschillende eigenschappen van een entiteit een verschillend discriminerend vermogen kunnen hebben, is dit niet noodzakelijk zo voor het samenvoegen van coreferente

objecten. Daarom zal een alternatieve methode worden bestudeerd. Meer specifiek zal er voor elk deelobject een samenvoegingsfunctie worden verondersteld en zal de samenvoeging van complexe objecten bestaan uit samenvoeging van de verschillende deelobjecten. Een dergelijke samenvoegingsfunctie noemen we samengesteld en is als volgt gedefinieerd.

Definitie 9.8 (Samengestelde samenvoegingsfunctie)
Laat $O = U_1 \times ... \times U_n$ een complex universum van objecten zijn. Een samengestelde samenvoegingsfunctie ϖ_O over O is gedefinieerd als:

$$\varpi_O : \mathcal{M}(O) \to O \tag{9.42}$$

waarbij:

$$\varpi_O(M) = \left(\varpi_{U_1}\left(\text{proj}_1\left(M\right)\right), ..., \varpi_{U_n}\left(\text{proj}_n\left(M\right)\right)\right) \tag{9.43}$$

en waarbij:

$$\text{proj}_i(M) \in \mathcal{M}(U_i) \tag{9.44}$$

zodat:

$$\omega_{\text{proj}_i(M)}(u) = \sum_{o \in M \wedge \text{proj}_i(o) = u} \omega_M(o). \tag{9.45}$$

Het kan dan makkelijk worden geverifieerd dat als ϖ_O wordt samengesteld uit idempotente functies ϖ_{U_i}, ϖ_O eveneens idempotent is. Als elke functie ϖ_{U_i} voldoet aan de meerderheidsregel, dan zal ook ϖ_O voldoen aan de meerderheidsregel. Als elke functie ϖ_{U_i} k-meerderheidsonafhankelijk is, dan zal ook ϖ_O dit zijn. Een eigenschap die niet wordt overgedragen, is bewaring. Inderdaad is het zo dat wanneer elke functie ϖ_{U_i} bewarend is, dit niet noodzakelijk geldt voor ϖ_O. Echter, we kunnen op basis van een samengestelde functie ϖ_O makkelijk een bewarende functie afleiden als volgt. Door de bewaring van ϖ_{U_i} kunnen we de bronnen traceren die aanleiding geven tot $\varpi_O(M)$. De betrouwbaarheid van een bron kan dan worden gemeten als het aantal bijdrages van een bron tot $\varpi_O(M)$ en een bewarende samenvoegingsfunctie wordt dan verkregen door keuze van de meest betrouwbare bron. Anders gezegd kunnen we een samenvoegingsfunctie construeren als volgt:

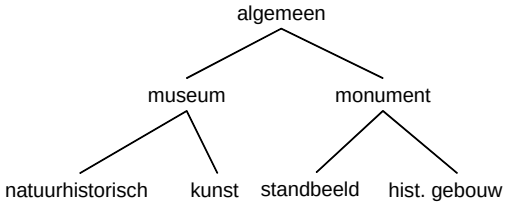
$$\varpi_O^* : \mathcal{M}(O) \to O \tag{9.46}$$

zodat:

$$\varpi_O^*\left(M\right) = \arg\max_{o \in M}\left(\omega_M(o) \cdot \left|\left\{i \middle| \, \text{proj}_i(o) = \text{proj}_i(\varpi_O(M))\right\}\right|\right) \tag{9.47}$$

waarbij ϖ_O een samengestelde samenvoegingsfunctie is. ϖ_O^* is dan een bewarende samenvoegingsfunctie geïnduceerd door ϖ_O.

Ter afsluiting bestuderen we voor complexe samenvoegingsfuncties het principe van λ-bewaring. Het nadeel van bewaring voor complexe objecten is namelijk dat de resulterende samenvoeging weinig robuust is tegen fouten. Wanneer door bewaring een object wordt gekozen, dat een goede beschrijving is van



Figuur 9.6: Relatie voor de eigenschap 'type'

het merendeel van de eigenschappen, dan bestaat de kans dat voor sommige eigenschappen een minder goede beschrijving wordt verkregen. Dit kan echter opgelost worden door bewaring te garanderen voor logische groepen van eigenschappen, d.i. groepen van eigenschappen bepaald door de groeperingsfunctie λ. Daarom definiëren we eerst een λ-partitie.

Definitie 9.9 (λ-partitie)
Laat $O = U_1 \times ... \times U_n$ een complex universum van objecten zijn met labelfunctie lab. *Een λ-partitie van de verzameling van labels is een partitie $\{P_j | j \in \{1, ..., k\}\}$ zodat er geldt dat:*

$$\forall j \in \{1, ..., k\} : \lambda(P_j) = 1. \tag{9.48}$$

Eigenschap 9.7 (λ-bewaring)
Laat $O = U_1 \times ... \times U_n$ een complex universum van objecten zijn. Een samenvoegingsfunctie ϖ_O is λ-bewarend ten opzichte van een λ-partitie $\{P_j | j \in \{1, ..., k\}\}$ als ϖ_O bewarend is ten opzichte van elke P_j.

In een extreem geval komt elk P_j uit de λ-partitie overeen met een label. Voor een dergelijke triviale partitie is een samengestelde samenvoegingsfunctie steeds λ-bewarend. De sterkte van λ-bewaring hangt dus af van de partitie die wordt beschouwd. Het voordeel van λ-bewaring is dat verschillende objecten kunnen bijdragen tot het samengevoegde object, met behoud van logische deelobjecten. Fouten in de resulterende objecten kunnen hierdoor beter worden vermeden.

Voorbeeld 9.4
Beschouwen we de objecten uit Tabel 9.1. Voor de eigenschap 'naam' veronderstellen we een samenvoegingsfunctie ϖ_S gebaseerd op E^*_S met parameters $(1, 0, 0.05)$. Laat ons veronderstellen dat de samenvoeging voor 'lengtegraad' en 'breedtegraad' bewarend is door een getal te kiezen dat zo dicht mogelijk bij de mediaan van de aanwezige deelobjecten ligt. Voor de eigenschap 'type' veronderstellen we de orderelatie R zoals getoond in Figuur 9.6. Wanneer we samenvoegingsfuncties toepassen op de verschillende deelobjecten verkrijgen we de complexe objecten getoond in Tabel 9.3.

Uit Tabel 9.3 kunnen we afleiden dat, hoewel alle samenvoegingsfuncties voor atomaire universa bewarend zijn, de samenvoegingsfunctie voor complexe

naam	lengtegraad	breedtegraad	type
Huis van Alijn	3.723372	51.057517	Hist. gebouw
Belfort en Lakenhalle	3.724911	51.053555	Hist. gebouw
SMAK	3.722726	51.038081	Museum

Tabel 9.3: Resultaat van samenvoeging met ϖ_O

objecten niet bewarend is. We kunnen op basis van deze samenvoegingsfunctie voor complexe objecten wel een bewarende samenvoegingsfunctie ϖ^*_O induceren. Als $\varpi_O(M)$ het resultaat is van samenvoeging van M, dan tellen we voor elk object uit M het aantal gemeenschappelijke deelobjecten met $\varpi_O(M)$. Voor het samengevoegde object met naam 'Huis van Alijn', bestond de multiverzameling M uit vier objecten (Tabel 9.1). Het eerste object heeft drie deelobjecten gemeenschappelijk met het samengevoegde, het tweede object heeft één deelobject gemeenschappelijk, het derde object heeft er twee gemeenschappelijk en het vierde object geen enkel. Dit betekent dat het eerste deelobject wordt gekozen. Passen we deze strategie toe op de drie multiverzamelingen uit Tabel 9.1, dan krijgen we als resultaat de objecten getoond in Tabel 9.4. Het kan worden geverifieerd dat deze samenvoeging inderdaad bewarend is.

naam	lengtegraad	breedtegraad	type
Huis van Alijn	3.723372	51.057411	Hist. gebouw
Belfort en Lakenhalle	3.724911	51.053653	Monument
SMAK	3.722726	51.038081	Algemeen

Tabel 9.4: Resultaat van samenvoeging met ϖ^*_O

We leiden uit Tabel 9.4 af dat het resultaat van de bewarende samenvoeging door ϖ^*_O enkele storende neveneffecten heeft. Het is bijvoorbeeld zo dat voor de eigenschap 'type' niet altijd de meest specifieke meting wordt gekozen, hetgeen een gevolg is van de bewaring van objecten. We kunnen echter gebruik maken van de groeperingsfunctie λ (zie ook Voorbeeld 2.3):

$$\begin{aligned}
\lambda(\{\mathit{lengtegraad}\}) &= 1\\
\lambda(\{\mathit{breedtegraad}\}) &= 1\\
\lambda(\{\mathit{naam}\}) &= 1\\
\lambda(\{\mathit{type}\}) &= 1\\
\lambda(\{\mathit{lengtegraad}, \mathit{breedtegraad}\}) &= 1\\
\lambda(\{\mathit{lengtegraad}, \mathit{breedtegraad}, \mathit{naam}, \mathit{type}\}) &= 1.
\end{aligned}$$

Deze functie leert ons dat de eigenschappen 'lengtegraad' en 'breedtegraad' samen een semantisch geheel vormen (Hoofdstuk 2), d.i. zij bepalen samen de geografische locatie van een POI. Beschouwen we nu de λ-partitie van de verzameling van labels $\{\{\mathit{naam}\}, \{\mathit{lengtegraad}, \mathit{breedtegraad}\}, \{\mathit{type}\}\}$, dan kunnen

we inzien dat de objecten getoond in Tabel 9.5 λ-bewarend zijn met betrekking tot deze partitie.

naam	lengtegraad	breedtegraad	type
Huis van Alijn	3.723372	51.057411	Hist. gebouw
Belfort en Lakenhalle	3.724911	51.053653	Hist. gebouw
SMAK	3.722726	51.038081	Museum

Tabel 9.5: Resultaat van λ-bewarende samenvoeging

9.6 Conclusie

In dit hoofdstuk is bestudeerd hoe een groep van coreferente objecten kan worden verwerkt. Meer specifiek is onderzocht hoe samenvoeging van coreferente objecten kan worden gedefinieerd door gebruik te maken van evaluatoren. Vertrekkende van een algemene definitie voor samenvoegingsfuncties zijn eigenschappen voor samenvoegingsfuncties opgesomd. Vervolgens is bestudeerd hoe samenvoegingsfuncties geconstrueerd kunnen worden op basis van een evaluator E_U. De eigenschappen van deze samenvoegingsfuncties zijn onderzocht. In het bijzonder is bestudeerd hoe semantische evaluatoren zich gedragen. Daaruit is gebleken dat het aanwezig zijn van een relatie beter rechtstreeks wordt gebruikt voor samenvoeging. Dit leidt tot de definitie van semantische samenvoeging. Ten slotte is onderzocht hoe samenvoegingsfuncties voor complexe objecten kunnen worden samengesteld uit samenvoegingsfuncties voor atomaire objecten. Ook de overdracht van eigenschappen is hierbij bestudeerd.

Hoofdstuk 10

Besluit en verder onderzoek

Ter afsluiting van deze thesis geven we een overzicht van de innovatieve bijdragen die zijn geleverd. Ook geven we enkele suggesties voor verder onderzoek.

10.1 Geleverde bijdragen

We zijn dit werk begonnen met het definiëren van objecten en entiteiten. Een object beschrijft een entiteit wat wordt gekarakteriseerd door een **referentiefunctie** ρ. Er is een onderscheid gemaakt tussen atomaire en complexe objecten en verschillende functies zijn gedefinieerd om de structuur van complexe objecten formeel te beschrijven. Een atomair object beschrijft een entiteit rechtstreeks. Een complex object bestaat uit een aantal deelobjecten die elk een welbepaalde eigenschap van een entiteit beschrijven. Wanneer twee objecten dezelfde entiteit beschrijven, spreken we van coreferente objecten. Het coreferent zijn van twee objecten wordt genoteerd als $o_1 \leftrightarrow o_2$. De taak waarbij, voor een gegeven verzameling van objecten, alle coreferente objecten moeten worden gezocht, wordt het coreferentieprobleem genoemd. Gelet op het feit dat coreferentie van objecten gedefinieerd is als gelijkheid van de beschreven entiteiten, weten we dat $\leftrightarrow$ een equivalentierelatie is. Door de creatie van objecten te benaderen als een (al dan niet fysisch) **meetproces**, stellen we vast dat imperfecties van dit meetproces leiden tot situaties waarbij verschillende objecten dezelfde entiteit beschrijven. We hebben een overzicht van dergelijke imperfecties gegeven en deze meetimperfecties zijn van die aard, dat het bepalen van coreferentie geen triviaal probleem is.

In dit werk is een **nieuwe possibilistische oplossing** gezocht voor het coreferentieprobleem. Meer bepaald, gelet op het feit dat $\leftrightarrow$ een equivalentierelatie is, is het coreferentieprobleem een probleem van Boolese aard. Hiermee wordt bedoeld dat twee objecten ofwel coreferent, ofwel niet coreferent zijn. De onzekerheid over coreferentie die wordt veroorzaakt door meetimperfecties, kan daarom worden gemodelleerd door een possibiliteitsverdeling over het domein van Boolese waarheidswaarden (d.i. $\mathbb{B} = \{T, F\}$). Een dergelijke possibiliteits-

verdeling wordt een possibilistische waarheidswaarde genoemd. Onze aanpak voor coreferentiebepaling vertrekt daarom van een operator die voor twee objecten een possibilistische waarheidswaarde construeert. Deze possibilistische waarheidswaarde geeft enerzijds de mogelijkheid dat twee objecten coreferent zijn en anderzijds de mogelijkheid dat twee objecten niet coreferent zijn. Een operator die deze possibilistische waarheidswaarde construeert voor twee objecten, wordt een **evaluator** genoemd. Gelet op het feit dat $\leftrightarrow$ een equivalentierelatie is (reflexief, symmetrisch en transitief), is onderzocht hoe deze **eigenschappen overgedragen kunnen worden naar een evaluator**. Het blijkt dat reflexiviteit en symmetrie onder redelijke veronderstellingen altijd van toepassing zijn op een evaluator. Transitiviteit is in de praktijk moeilijk te garanderen. Naast het coreferentieprobleem zijn twee verwante problemen aangehaald, die beiden onder de noemer '**gedeeltelijke coreferentie**' vallen. Het is doorheen deze thesis aangetoond hoe deze problemen kunnen worden opgelost met de nieuwe methoden die zijn voorgesteld als oplossing voor het coreferentieprobleem.

Na studie van de oorzaken van onzekerheid bij coreferentie is gebleken dat sommige van deze oorzaken algemeen kunnen worden aangepakt. Hiermee wordt bedoeld dat het omgaan met deze onzekerheid, losstaat van het universum waarin objecten worden uitgedrukt. Een eerste voorbeeld hiervan is de onzekerheid die wordt veroorzaakt door niet-meetbare en moeilijk-meetbare eigenschappen. Om deze te behandelen, construeren we een **evaluator voor possibilistische variabelen** over O, die gebaseerd is op een evaluator over O. Een tweede voorbeeld is dat van **semantische evaluatie**, waarbij een binaire relatie evidentie biedt voor het coreferent zijn van twee objecten. In onze nieuwe aanpak voor semantische evaluatie onderzoeken we hoe deze kennis kan worden gebruikt voor de generatie van possibilistische waarheidswaarden. Naast de klasse van semantische evaluatoren, bestaat ook de klasse van **syntactische evaluatoren**. Deze evaluatoren baseren zich op een syntactische vergelijking van twee objecten voor de constructie van een possibilistische waarheidswaarde.

In onze zoektocht naar nieuwe syntactische evaluatoren ondervinden we de noodzaak om possibilistische waarheidswaarden te combineren. Om die reden is onderzoek gedaan naar combinatiefuncties. We gaan hierbij uit van één Boolese propositie p en n actoren die elk op zich onzekerheid formuleren over de waarheidswaarde van p. We willen dan een functie construeren die al deze onzekerheid combineert in één possibilistische waarheidswaarde. Een overzicht van de literatuur over combinatiefuncties leert ons dat de uitbreidingsprincipes van Zadeh en De Cooman toelaten om de Boolese functies $\wedge$ en $\vee$ uit te breiden naar het domein van possibilistische waarheidswaarden. De Cooman modelleert hierbij expliciet afhankelijkheden die tussen actoren kunnen bestaan. We hebben geargumenteerd dat de uitbreiding van klassieke Boolese functies niet voldoende is om het coreferentieprobleem aan te pakken. De Tré en De Baets beschrijven een methode voor transformatie van possibilistische waarheidswaarden alvorens deze te combineren. Hoewel deze laatste methode enkele gewenste eigenschappen bezit, is aangetoond dat de methode van De Tré en De Baets

enkele nadelen heeft, zoals bijvoorbeeld een beperkte beeldverzameling. Onze nieuwe aanpak voor combinatie van onzekerheid vertrekt van afhankelijkheden tussen actoren die orthogonaal staan op afhankelijkheden in de zin van De Cooman. Dergelijke afhankelijkheden kunnen worden gemodelleerd door het vertrouwen in groepen van actoren in rekening te brengen. Dit vertrouwen modelleren we aan de hand van twee vertrouwensmaten die we samen de **conditionele necessiteit** noemen. De eigenschappen van conditionele necessiteit leren ons dat de structuur ervan aan strenge regels onderhevig is. Na de studie van conditionele necessiteit bepalen we een methode voor de combinatie van enerzijds de kennis P_π en anderzijds de conditionele necessiteit. Dit leidt tot een discrete **Sugeno-integraal voor possibilistische waarheidswaarden**. De **eigenschappen** van deze nieuwe combinatiefunctie zijn onderzocht en het is beschreven hoe ze zich verhoudt tot enerzijds de methode van De Tré en De Baets en anderzijds de methode van De Cooman. Enkele bijzondere gevallen van conditionele necessiteit leiden tot **bijzondere gevallen van de Sugeno-integraal**. Zo blijkt dat de uitersten van de Sugeno-integraal gelijk zijn aan de Zadeh-uitbreiding van $\wedge$ en $\vee$.

De nieuwe combinatiefunctie voor possibilistische waarheidswaarden wordt vervolgens gebruikt bij de constructie van syntactische **evaluatoren voor verzamelingen en multiverzamelingen** van objecten uit een universum O. Hiervoor wordt een nieuw algoritme voorgesteld dat een **leximax-optimale één-op-één afbeelding** ι construeert tussen twee (multi)verzamelingen van objecten. Na een complexiteitsanalyse worden suggesties gedaan om de uitvoeringstijd te versnellen. Door een evaluator over O te gebruiken, leidt de afbeelding ι tot een sequentie van possibilistische waarheidswaarden. Deze sequentie kan worden gecombineerd met de nieuwe Sugeno-integraal voor possibilistische waarheidswaarden, op voorwaarde dat de gebruikte conditionele necessiteit aan **welbepaalde voorwaarden** voldoet. Onder deze voorwaarden kan de conditionele necessiteit worden voorgesteld als een gewichtsvector en kan de **Sugeno-integraal worden herschreven** in termen van de Zadeh-uitbreiding van $\wedge$. Voor de generatie van de gewichtsvector is een **nieuwe geparameteriseerde functie** voorgesteld die aan interessante eigenschappen voldoet. Er wordt aangetoond dat de evaluator voor collecties slechts transitief is als de combinatiefunctie gelijk is aan de Zadeh-uitbreiding van $\wedge$ en de evaluator over O transitief is. Tot slot wordt het geval van **meervoudige kwantificatie** besproken.

De nieuwe evaluator voor (multi)verzamelingen wordt vervolgens gebruikt bij de constructie van nieuwe syntactische evaluatoren voor strings. Hiervoor wordt eerst een studie gemaakt van operatoren voor strings. Op basis daarvan wordt een **één-niveau evaluator** gedefinieerd. Deze één-niveau evaluator heeft twee belangrijke eigenschappen. Enerzijds heeft de evaluator een **lage complexiteit**, anderzijds is de toekenning van possibiliteiten gebaseerd op een taxonomie van deelverschillen tussen strings, zodat **verschillende soorten van fouten verschillend beoordeeld** kunnen worden. We hebben enkele voorbeelden gegeven van toekenningsmodellen. De één-niveau evaluator wordt

vervolgens gebruikt bij de constructie van een **twee-niveau evaluator**. Hierbij wordt een splitsingsfunctie gebruikt om een string om te zetten naar een multiverzameling van strings. Deze multiverzamelingen kunnen worden vergeleken met behulp van de eerder gedefinieerde vergelijkingsoperator voor multiverzamelingen. De voorgestelde geparameteriseerde functie voor de bepaling van een gewichtsvector wordt onderzocht in een praktische context. Dit leidt tot nieuwe methoden voor de bepaling van de parametervector, enerzijds met en anderzijds zonder trainingsdata. Er wordt een schema gegeven voor de iteratieve bepaling van possibiliteit. Het is aangetoond dat in een bijzonder geval, deze iteratieve bepaling leidt tot een filter met lage complexiteit. Hierdoor wordt er voor veel koppels van strings vermeden dat de evaluator moet worden uitgevoerd, hetgeen een optimalisatie betekent. Het is gebleken uit experimenten dat de filter slechts een beperkte invloed heeft op de accuraatheid van de tweeniveau evaluator. Uit dezelfde experimenten is ook gebleken dat onze nieuwe aanpak meestal beter presteert dan de best presterende vergelijkingstechniek voor strings uit de literatuur (SoftTFIDF).

Er is onderzoek gedaan naar het oplossen van de **inconsistenties** die kunnen voorkomen bij evaluatoren. Een eerste probleem dat we hebben bestudeerd, is het **niet-transitief zijn van evaluatoren**. Om dit probleem aan te pakken zijn **beslissingsmodellen** ingevoerd. Tweewaardige beslissingsmodellen drukken een beslissing uit in het Boolese domein op basis van een possibilistische waarheidswaarde. Driewaardige beslissingsmodellen drukken een beslissing uit in de machtsverzameling van het Boolese domein op basis van een possibilistische waarheidswaarde. Een driewaardig beslissingsmodel is nuttig wanneer het nemen van een foute beslissing een te hoge kost heeft, zodat het beter is expliciet aan te geven dat een beslissing niet kan worden genomen. Op basis van deze beslissingsmodellen kunnen we relaties afleiden die een schatting vormen van de coreferentierelatie $\leftrightarrow$. Als de afgeleide relatie transitief is, dan kan ze worden gebruikt voor de generatie van een partitie van de objectruimte. Partitieklassen van deze partitie bevatten dan objecten die onderling coreferent zijn. Onze aanpak om de problematiek rond niet-transitieve evaluatoren op te lossen, speelt hierop in door op basis van de niet-transitieve relatie een partitie te construeren. Aangezien een dergelijke partitie overeenkomt met een equivalentierelatie, verkrijgen we op die manier een benadering van $\leftrightarrow$. Voor het genereren van de partitie tonen we het opvallende verband aan tussen het hiërarchische clusteralgoritme en het raamwerk van possibiliteiten. Dit geeft aanleiding tot een **Sugeno-gebaseerde methode** voor het herstel van transitiviteit. Als alternatief bespreken we ook de generatie van een partitie door **maximale en voldoende klassen** af te splitsen. Een tweede probleem dat we hebben bestudeerd, is het eventueel aanwezig zijn van **randvoorwaarden**. Wanneer twee databanken op zich geen coreferente objecten bevatten[1], dan weten we dat elk object uit de ene databank ten hoogste met één object uit de andere databank coreferent is. Deze randvoorwaarde heeft toepassingen in

[1]Een object is uiteraard coreferent met zichzelf. We bedoelen hier coreferente objecten die verschillen van elkaar.

onder andere identificatieproblemen en websitevergelijking. Om deze randvoorwaarde in rekening te brengen kan het algoritme voor de constructie van een afbeelding ι in de context van vergelijking van (multi)verzamelingen worden toegepast.

Na de studie van evaluatoren voor atomaire objecten is onderzoek gedaan naar **evaluatoren voor complexe objecten**. Hierbij wordt voor elk deeluniversum van het complexe universum een deelevaluator voorzien. Vervolgens worden eventuele parameters van deze deelevaluatoren bepaald. Elke deelevaluator maakt daarbij enkel gebruik van het overeenkomstige deeluniversum, zodat de parameters van deelevaluatoren (indien nodig) in parallel kunnen worden bepaald. We hebben aangetoond hoe een binaire relatie voor semantische evaluatoren dynamisch kan worden opgebouwd door te steunen op de verzamelde kennis van syntactische evaluatoren. Vervolgens is bestudeerd hoe conditionele necessiteit kan worden geconstrueerd. Hierbij tonen we aan hoe het gebruik van beslissingsmodellen enerzijds en de wet van ononderscheidbaarheid anderzijds leidt tot sterke beperkingen van de vertrouwensmaten. Deze beperkingen hebben als gevolg dat het leerproces sterk wordt vereenvoudigd. Onze aanpak voor de bepaling van de vertrouwensmaten is gebaseerd op een schatting van de *zuiverheid*. Experimentele resultaten tonen aan dat onze aanpak, ondanks de sterke beperkingen op de vertrouwensmaten, beter presteert dan de probabilistische methode van Fellegi en Sunter en het k-means clusteralgoritme.

Tot hier toe is bij het onderzoek van evaluatoren voor atomaire en complexe objecten uitgegaan van een goed gestructureerde databank waarin objecten worden opgeslagen. Deze veronderstelling is niet van toepassing in het geval van tekstuele beschrijvingen van entiteiten. Het probleem van coreferente tekstuele beschrijvingen staat in de literatuur bekend als tekstclustering. In onze aanpak van coreferentiebepaling van teksten vertrekken we van een fundamenteel **nieuw tekstmodel gebaseerd op multirelaties** over een conceptruimte $\mathcal{C}$. Voor de transformatie van tekst naar deze voorstellingsruimte wordt gebruik gemaakt van de i.Know-technologie. Deze technologie laat toe om dynamisch concepten te extraheren uit teksten, zonder een vooraf bepaalde ontologie. Concepten komen overeen met een semantisch geheel en kunnen door de i.Know technologie efficiënt worden gevonden, los van hun lengte en complexiteit. Bovendien genereert de i.Know-technologie ook de bestaande relaties tussen de gevonden concepten. Voor de bepaling van coreferentie van twee teksten is een bijkomend meetproces ingevoerd, dat voor een gegeven tekst de relevante koppels van concepten aanduidt. Een probleem hierbij is het onbekend zijn van dit meetproces. Dit probleem lossen we op door veronderstellingen te maken over enerzijds de evaluator voor teksten en anderzijds het onbekende meetproces. Hieruit blijkt dat we efficiënt te werk kunnen gaan door rechtstreeks een partitie over de tekstruimte te genereren. We hebben een methode gegeven voor een **optimale bepaling van een dergelijke partitie**. We hebben aangetoond hoe het **aantal clusters kan worden bepaald** op basis van de veronderstelling van de Wet van Zipf. Met het nieuwe tekstmodel kunnen we een verzameling van documenten voorstellen als één multirelatie.

Voor elke k-snede van deze multirelatie kunnen we clusters genereren zodat de *zuiverheid* van de clusters hoog is. We kunnen het aantal gegenereerde clusters vergelijken met het geschat aantal clusters, wat ons een controle van de *compleetheid* oplevert. Op die manier komen we tot een **optimale balans** tussen *zuiverheid* en *compleetheid*. Experimentele resultaten tonen aan dat onze aanpak een **zeer grote verbetering** biedt ten opzichte van standaardmethoden voor clustering. Ten opzichte van geavanceerde methoden zoals Latente Dirichlet Allocatie (LDA) blijkt dat onze aanpak het minder doet op gebied van *compleetheid*, maar veel beter op gebied van *zuiverheid*. In het kader van tekstsamenvatting is een hoge *zuiverheid* echter belangrijker dan een hoge *compleetheid*. Een belangrijke vaststelling is dat de lage *compleetheid* te wijten is aan het niet aanwezig zijn van relevante patronen in bepaalde documenten. Dit betekent dat onze aanpak voor clusteren van teksten kan worden verbeterd door bijkomende clustervorming op basis van bijvoorbeeld conceptanalyse.

Tot slot van deze thesis is onderzocht hoe coreferente objecten kunnen worden samengevoegd tot één object dat een zo goed mogelijke beschrijving geeft van de gerefereerde entiteit. Hiervoor zijn **samenvoegingsfuncties** gebaseerd op een evaluator gedefinieerd en zijn een aantal **relevante eigenschappen** opgesomd. Een belangrijke eigenschap is bewaring. Dit betekent dat de samenvoeging van een multiverzameling van objecten een element is van de gegeven multiverzameling. Er is aangetoond hoe samenvoegingsfuncties voor een atomaire universum geconstrueerd kunnen worden op basis van een evaluator over dit universum. Voor een gegeven multiverzameling M van samen te voegen objecten, wordt voor elk object u in M een vaag natuurlijk getal opgebouwd. Dit natuurlijk getal modelleert het aantal objecten in M waarmee u coreferent is, rekening houdend met de onzekerheid gespecificeerd door de evaluator. Op die manier kunnen we objecten uit M ordenen aan de hand van een ordening van vage natuurlijke getallen. Er is onderzocht welke eigenschappen de ingevoerde klasse van samenvoegingsfuncties bezit. In het kader van de samenvoeging van (multi)verzamelingen is geargumenteerd waarom bewaring een te strenge eigenschap is. Vervolgens is bestudeerd hoe samenvoegingsfuncties voor complexe objecten kunnen worden **samengesteld** op basis van samenvoegingsfuncties voor deelobjecten. Hierbij is aandacht besteed aan de overdracht van eigenschappen. Ook in het kader van complexe objecten is gebleken dat bewaring een te strenge eigenschap is. Als alternatieve eigenschap is **λ-bewaring** voorgesteld.

10.2 Verder onderzoek

Om te komen tot een praktisch instrument voor automatische detectie en verwerking van coreferente objecten, moeten een aantal zaken (verder) worden onderzocht. Om die reden geven we hier enkele voorstellen voor verder onderzoek.

Ten eerste ligt binnen deze thesis de nadruk op data van tekstuele aard. In het geval van complexe objecten hebben we steeds strings als deelobjecten

verondersteld en een belangrijk deel van dit werk is gericht op coreferentie van tekstuele beschrijvingen. Evaluatoren voor andere universa zijn meestal vrij eenvoudig te construeren. Voor numerieke universa zoals $\mathbb{R}$ en $\mathbb{N}$ kunnen evaluatoren gebaseerd zijn op afstandsmaten. Echter, een cruciaal punt in onze aanpak van coreferentie van complexe objecten is het bepalen van de parameters van deelevaluatoren. We hebben aangehaald in Hoofdstuk 6 dat de schattingsmethode voor strings ook toepasbaar is op numerieke data. Om deze bewering voldoende hard te maken is echter verder onderzoek nodig.

Ten tweede achten we het nuttig verder onderzoek te doen naar het belang van meetprocessen en meer bepaald de imperfecties ervan. Bij meting door transformatie in de context van beelden, kan de kwaliteit van beelden een ernstige beperking vormen op de kwaliteit van de meting. Echter, er bestaan heel wat situaties waarbij het verzamelen van meerdere coreferente beelden een lage kost heeft. Als al deze beelden een lage kwaliteit hebben, zullen de objecten na meting ook een lage kwaliteit hebben (d.i., de meting van entiteiten is onderhevig aan zware imperfecties). Echter, door gebruik te maken van de kennis dat deze objecten coreferent zijn, is het mogelijk één object van een hogere kwaliteit te construeren. Hiervoor kunnen samenvoegingsfuncties een uitgangspunt vormen, maar er zijn heel wat andere mogelijkheden die te onderzoeken zijn. Aansluitend bij de slechte kwaliteit van meetprocessen kan het interessant zijn om driewaardige beslissingsmodellen te gebruiken. Het niet nemen van een beslissing kan dan voortkomen uit de kennis dat de meting onbetrouwbaar is.

Ten derde stellen we vast dat nog heel wat onderzoek kan worden gedaan naar coreferentiebepaling en samenvoeging van teksten. Voor coreferentiebepaling is in Hoofdstuk 8 aangehaald wat de grenzen van onze aanpak zijn. Er is geargumenteerd dat een studie van concepten kan helpen bij het verbeteren van de nieuwe methode. Laat ons dit nog even verder toelichten. Wanneer een tekst afzonderlijk wordt geobserveerd, is het bijzonder moeilijk om automatisch te bepalen welke concepten of patronen kenmerkend zijn voor het onderwerp dat die tekst beschrijft. Dit is de reden waarom het bijkomende meetproces in Hoofdstuk 8 onbekend is. Echter, wanneer onze aanpak wordt toegepast, verkrijgen we clusters met hoge zuiverheid. Dit betekent dat teksten in deze clusters met hoge zekerheid onderling coreferent zijn. Gelet op de eigenschap dat de relevantie van concepten en patronen wordt bepaald door hun multipliciteit, betekent dit dat we in deze clusters op zoek kunnen gaan naar relevante concepten. Deze concepten kunnen dan worden gebruikt om clusters verder aan te vullen met teksten die door onze aanpak in singleton clusters zijn geplaatst. Een zo mogelijk nog grotere uitdaging is de samenvoeging van coreferente teksten. Dit probleem staat in de literatuur bekend als *meervoudige documentsamenvatting* (MDS). De literatuur hierover is echter uiterst beperkt. Daarom pleiten we hier voor een formele studie van dit probleem, vertrekkende vanuit het raamwerk van samenvoegingsfuncties. In het bijzonder moet hierbij aandacht worden besteed aan eigenschappen verwant aan bewaring. Een vertrekpunt voor dergelijke verwante eigenschappen zijn *maximaal coherente*

deelverzamelingen, die in de context van possibilistische samenvoeging zijn bestudeerd door De Stercke en Dubois [143].

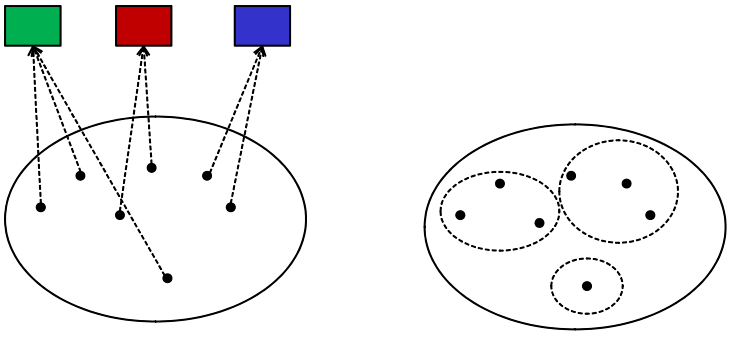
Bijlage A

Clustertechnieken

A.1 Het clusteringsprobleem

In deze bijlage wordt een kort overzicht gegeven van enkele clusteralgoritmen. Deze bijlage dient ter ondersteuning van de experimenten gerapporteerd in Hoofdstukken 7 en 8.

Clusteren is een probleem uit de *datamining* waarbij een groep van objecten moet worden gepartitioneerd. Anders dan bij classificatie zijn de klassen niet vooraf gespecificeerd. Meestal wordt het aantal klassen als gekend verondersteld, hoewel deze veronderstelling in de praktijk niet altijd opgaat. Het verschil tussen clustering en classificatie wordt geïllustreerd in Figuur A.1.



Figuur A.1: Classificatie (links) versus clusteren (rechts)

A.2 Vectorruimten

Een veel gebruikte aanpak voor de voorstelling van data is het vectorruimtemodel. Hierbij wordt elk object voorgesteld als een vector $\mathbf{v}$ met m numerieke

dimensies. Wanneer n objecten gegeven zijn, wordt de datacollectie voorgesteld als een $n \times m$-matrix $\mathbf{M}$. Elke rij van deze matrix komt overeen met één object en elke kolom komt overeen met één dimensie. In deze bijlage noteren we de waarde van de j^{de} dimensie voor het i^{de} object als $\mathbf{v}_j^i$ of als $\mathbf{M}(i,j)$.

A.3 Principale Componenten Analyse

In sommige toepassingen kan het aantal dimensies m zeer hoog oplopen. Dit is bijvoorbeeld het geval bij het voorstellen van teksten in een vectorruimtemodel. Een dergelijke hoge dimensionaliteit heeft als gevolg dat het clusteren heel wat complexer wordt. Daarnaast heeft een hoge dimensionaliteit ook een negatief effect op de accuraatheid van het clusteren. Om die reden wordt voor data met een groot aantal dimensies meestal een dimensiereductie doorgevoerd. Een veelgebruikte methode om dit te doen is Principale Componenten Analyse (PCA). Deze methode reduceert het aantal dimensies door enkel die (combinaties van) dimensies over te houden, die de meeste variantie verklaren. Dimensiereductie met PCA is een proces bestaande uit vijf stappen, die we hier één voor één kort bespreken.

In de eerste stap wordt de data genormaliseerd langs de dimensies. Hiervoor wordt voor elke dimensie het gemiddelde van die dimensie afgetrokken. Noteren we:

$$\forall j \in \{1, ..., m\} : \overline{x}_j = \frac{1}{n} \sum_{i=1}^{n} \mathbf{M}(i,j). \tag{A.1}$$

dan is de genormaliseerde matrix $\mathbf{M}^{\text{norm}}$ bepaald als:

$$\mathbf{M}^{\text{norm}}(i,j) = \mathbf{M}(i,j) - \overline{x}_j. \tag{A.2}$$

In de tweede stap wordt voor een datacollectie $\mathbf{M}$ de covariantiematrix van $\mathbf{M}^{\text{norm}}$ berekend. Als $\mathbf{M}$ een $n \times m$-matrix is, dan is de covariantiematrix $\text{cov}(\mathbf{M}^{\text{norm}})$ een $m \times m$-matrix waarvoor er geldt dat:

$$\text{cov}(\mathbf{M}^{\text{norm}})(k,l) = \frac{\sum_{i=1}^{n} (\mathbf{M}^{\text{norm}}(i,k)\mathbf{M}^{\text{norm}}(i,l))}{n-1}. \tag{A.3}$$

In de derde stap worden de eigenwaarden en bijhorende eigenvectoren berekend. Dit levert een m-dimensionale vector van eigenwaarden en een bijhorende $m \times m$-matrix waarin m vectoren van dimensie m worden voorgesteld. Gegeven een vierkante $m \times m$-matrix $\mathbf{A}$, een m-dimensionale vector $\mathbf{v}$ en een reëel getal e. Als er geldt dat:

$$\mathbf{A}\mathbf{v} = e\mathbf{v} \tag{A.4}$$

dan is $\mathbf{v}$ een eigenvector van $\mathbf{A}$ en e een eigenwaarde van $\mathbf{A}$. Wanneer $\mathbf{A}$ een lineaire transformatie voorstelt, dan is een eigenvector onder deze transformatie gelijk aan een schaling van zichzelf met e. We noteren de matrix $\mathbf{E}$ als de $m \times m$-matrix die op elke rij één van de m eigenvectoren bevat.

In een vierde stap worden de eigenwaarden en eigenvectoren gebruikt voor de reductie van de dimensies. De absolute waarde van een eigenwaarde is namelijk een maat voor de variantie die wordt verklaard door de overeenkomstige eigenvector. De eigenvector die overeenkomt met de grootste eigenwaarde wordt de *principale component* of hoofdcomponent genoemd. De reductie van dimensies wordt bereikt door een selectie te maken van de belangrijkste eigenvectoren (d.z. eigenvectoren met de grootste overeenkomstige eigenwaarden). Hiervoor bestaan verschillende mogelijkheden. Soms worden enkel de eigenvectoren gekozen waarvoor de overeenkomstige eigenwaarden een absolute waarde groter dan 1 hebben. In Hoofdstuk 8 zijn die eigenvectoren gekozen, zodat de som van de overeenkomstige eigenwaarden groter is dan 95% van de som van alle eigenwaarden. We noteren de matrix met de geselecteerde eigenvectoren als $\mathbf{E}^{\text{red}}$.

In de laatste en vijfde stap wordt een reconstructie van de datacollectie gemaakt op basis van de gekozen eigenvectoren. Dit gebeurt door toepassing van de geselecteerde eigenvectoren op de genormaliseerde datacollectie, na transponering.

$$\mathbf{M}^{\text{red}} = \mathbf{E}^{\text{red}} \times \left((\mathbf{M}^{\text{norm}})^T \right) \tag{A.5}$$

Merk op dat $\mathbf{E}^{\text{red}}$ m kolommen heeft en dat de getransponeerde van $\mathbf{M}^{\text{norm}}$ m rijen heeft, zodat de matrixvermenigvuldiging mogelijk is. In deze $\mathbf{M}^{\text{red}}$ bevinden de dimensies zich nu op de rijen, zodat een bijkomende transponering nodig is om compatibel te zijn met het vectorruimtemodel.

A.4 Het k-means algoritme

Een basismethode voor clustering is het *k-means* clusteralgoritme. Dit algoritme is gebaseerd op een afstandsmaat d voor vectoren. Veronderstellen we voor de eenvoud dat de vectorruimte waarin we werken gelijk is aan $\mathbb{R}^m$ en dat afstand wordt uitgedrukt als een reëel getal. In dat geval is een afstandsmaat gedefinieerd als een functie:

$$d : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R} \tag{A.6}$$

die moet voldoen aan de volgende drie voorwaarden:

$$\forall \mathbf{v} \in \mathbb{R}^m \quad : \quad d(\mathbf{v}, \mathbf{v}) = 0 \tag{A.7}$$

$$\forall (\mathbf{v}, \mathbf{w}) \in (\mathbb{R}^m)^2 \quad : \quad d(\mathbf{v}, \mathbf{w}) = d(\mathbf{w}, \mathbf{v}) \tag{A.8}$$

$$\forall (\mathbf{u}, \mathbf{v}, \mathbf{w}) \in (\mathbb{R}^m)^3 \quad : \quad d(\mathbf{u}, \mathbf{v}) + d(\mathbf{v}, \mathbf{w}) \geq d(\mathbf{u}, \mathbf{w}). \tag{A.9}$$

Een voorbeeld van een afstandsfunctie is de Minkowski-afstand:

$$d_p(\mathbf{v}, \mathbf{w}) = \left(\sum_{i=1}^{m} (\mathbf{v}_i - \mathbf{w}_i)^p \right)^{\frac{1}{p}} \tag{A.10}$$

die voor $p = 1$ gelijk is aan de Manhattan-afstand en die voor $p = 2$ gelijk is aan de Euclidische afstand.

Bij het *k-means* algoritme wordt er verondersteld dat het aantal clusters gekend is. Dit aantal clusters wordt genoteerd als k. De clusters worden genoteerd als:

$$C = \{C_1, ..., C_k\} \tag{A.11}$$

waarbij elke cluster een aantal vectoren uit de datacollectie kan bevatten. Voor elk van de clusters is de centroïde $\mathbf{m}$ gedefinieerd als de vector die in elke dimensie het gemiddelde van de dimensies van de aanwezige vectoren bevat. Er geldt bijgevolg:

$$\forall i \in \{1, ..., k\} : \mathbf{m}^{(i)} = \frac{1}{|C_i|} \sum_{\mathbf{v} \in C_i} \mathbf{v}. \tag{A.12}$$

Het doel van het *k-means* algoritme is de minimalisatie van de fout binnen clusters. Dit betekent dat voor elke cluster, de afstand tussen elke vector in die cluster en de centroïde zo klein mogelijk moet zijn. Dit kunnen we noteren als volgt:

$$C = \arg\min_{C} \sum_{i=1}^{k} \left(\sum_{\mathbf{v} \in C_i} d\left(\mathbf{v}, \mathbf{m}^{(i)}\right) \right). \tag{A.13}$$

Het algoritme om deze optimalisatie te bekomen, is een iteratieve uitvoering van twee stappen: de toekenningsstap en de updatestap.

In de toekenningsstap wordt elke vector uit de datacollectie toegekend aan één van de clusters. In iteratie t wordt de toekenning gedaan als volgt.

$$C_i^{(t)} = \left\{ \mathbf{v} | \forall j \in \{1, ..., k\} : d(\mathbf{v}, \mathbf{m}^{(i),(t)}) \leq d(\mathbf{v}, \mathbf{m}^{(j),(t)}) \right\}. \tag{A.14}$$

Dit betekent dat een vector wordt toegekend aan de cluster waarvoor de afstand tot de centroïde van de cluster, minimaal is. De eerste toekenning (d.i. $t = 0$) gebeurt willekeurig, vermits de clusters leeg zijn en er nog geen centroïdes berekend kunnen worden.

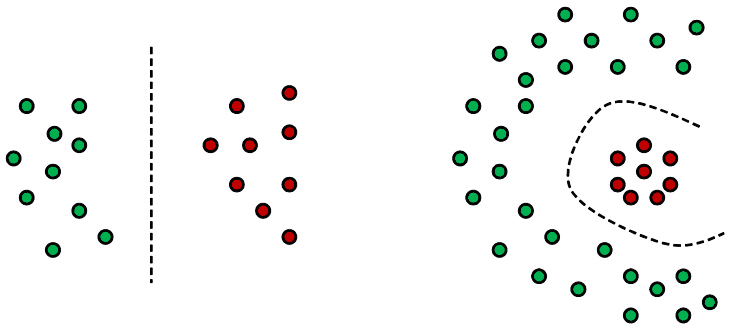
In de updatestap worden de centroïdes herberekend als volgt:

$$\mathbf{m}^{(i),(t+1)} = \frac{1}{|C_i^{(t)}|} \sum_{\mathbf{v} \in C_i^{(t)}} \mathbf{v}. \tag{A.15}$$

De nieuwe centroïdes worden in een volgende iteratie gebruikt om de toekenning van vectoren aan clusters te herberekenen.

Er bestaan verschillende variaties van dit algoritme met betrekking tot het stopcriterium. Een eerste mogelijkheid is stoppen wanneer het resultaat van twee opeenvolgende toekenningsstappen identiek is. In dat geval spreken we van stabiele clusters. Een tweede mogelijkheid is de specificatie van een drempelwaarde voor de clusterfout, gegeven door (A.13). Een derde mogelijkheid is de specificatie van een vast aantal iteraties.

Een eigenschap van het *k-means* clusteralgoritme is dat steeds lineair scheidbare clusters worden gevormd. Dit is in de praktijk niet altijd gewenst. Figuur A.2 illustreert het onderscheid tussen lineair en niet-lineair scheidbare clusters.



Figuur A.2: Lineair scheidbare clusters (links) versus niet-lineair scheidbare clusters (rechts)

Een mogelijke oplossing voor dit probleem is het beschouwen van een niet-lineaire afbeelding Φ naar een nieuwe ruimte. Toepassing van het *k-means* algoritme in deze nieuwe ruimte levert dan clusters die in de oorspronkelijke ruimte niet-lineair scheidbaar zijn. Dit betekent dat (A.13) wordt vervangen door:

$$C = \arg\min_{C} \sum_{i=1}^{k} \left(\sum_{\mathbf{v} \in C_i} d\left(\Phi(\mathbf{v}), \mathbf{m}^{(i)}\right) \right) \tag{A.16}$$

waarbij er geldt dat:

$$\forall i \in \{1, ..., k\} : \mathbf{m}^{(i)} = \frac{1}{|C_i|} \sum_{\mathbf{v} \in C_i} \Phi(\mathbf{v}). \tag{A.17}$$

Een probleem dat zich nu stelt, is dat de afbeelding Φ niet gekend is. Laat ons veronderstellen dat de afstandsmaat d de Euclidische afstand is. In dat geval geldt er dat:

$$d\left(\Phi(\mathbf{v}), \mathbf{m}^{(i)}\right) = \left\| \Phi(\mathbf{v}) - \mathbf{m}^{(i)} \right\|. \tag{A.18}$$

Het kwadraat van deze afstand kunnen we als volgt herschrijven:

$$\Phi(\mathbf{v}) \cdot \Phi(\mathbf{v}) - \frac{2}{|C_i|} \sum_{\mathbf{w} \in C_i} (\Phi(\mathbf{v}) \cdot \Phi(\mathbf{w})) + \frac{1}{|C_i|^2} \sum_{(\mathbf{u}, \mathbf{w}) \in C_i^2} (\Phi(\mathbf{u}) \cdot \Phi(\mathbf{w})). \tag{A.19}$$

Dit betekent dat we de afbeelding Φ als dusdanig niet hoeven te kennen, maar wel het resultaat van het product van twee vectoren die getransformeerd zijn met Φ. Hiervoor wordt gebruik gemaakt van een kernelfunctie κ, zodat er geldt dat:

$$\Phi(\mathbf{v}) \cdot \Phi(\mathbf{w}) = \kappa(\mathbf{v}, \mathbf{w}). \tag{A.20}$$

Veelterm	$\kappa(\mathbf{v}, \mathbf{w}) = (\mathbf{v} \cdot \mathbf{w} + x)^y$
Gauss	$\kappa(\mathbf{v}, \mathbf{w}) = \exp\left(\frac{-\|\mathbf{v}-\mathbf{w}\|^2}{2\sigma^2}\right)$
Sigmoïde	$\kappa(\mathbf{v}, \mathbf{w}) = \tanh(f(\mathbf{v} \cdot \mathbf{w}) + x)$
Cosinus	$\kappa(\mathbf{v}, \mathbf{w}) = \frac{\mathbf{v} \cdot \mathbf{w}}{\|\mathbf{v}\|\|\mathbf{w}\|}$

Tabel A.1: Voorbeelden van kernelfuncties

Enkele veelgebruikte kernelfuncties worden getoond in Tabel A.1. In Hoofdstuk 8 is gebruik gemaakt van de cosinus kernelfunctie. De variatie van het *k-means* algoritme dat gebruik maakt van een kernelfunctie wordt het kernel *k-means* algoritme genoemd.

A.5 Het hiërarchisch algoritme

Een tweede basismethode uit de literatuur is het hiërarchisch clusteralgoritme. Hierbij wordt een binaire boom opgebouwd waarvan de bladknopen overeenkomen met vectoren uit de datacollectie. Het opklimmen in de boom komt dan overeen met het samenbrengen van vectoren in één cluster. Strikt genomen bestaan er twee varianten van dit algoritme. De eerste variant vertrekt van één cluster met daarin alle vectoren uit de datacollectie. Deze cluster wordt dan iteratief opgesplitst. De tweede variant vertrekt van evenveel clusters als er vectoren zijn, zodat elke cluster precies één vector uit de datacollectie bevat. Deze clusters worden iteratief samengevoegd. In wat volgt bespreken we enkel de tweede variant.

Het hiërarchisch algoritme vertrekt net als het *k-means* algoritme van een afstandsmaat voor vectoren. Initieel zijn er evenveel clusters als vectoren in de datacollectie en bevat elke cluster precies één vector.

$$C = \{C_1, ..., C_n\}. \tag{A.21}$$

Vervolgens worden twee clusters gekozen als volgt:

$$(C_i, C_j) = \arg\min_{i,j} f_k(C_i, C_j). \tag{A.22}$$

Hierbij is f_k een kostfunctie. In Hoofdstuk 8 zijn de volgende twee kostfuncties gebruikt. De *enkelvoudige regel* schrijft voor dat:

$$f_k(C_i, C_j) = \min_{\mathbf{v} \in C_i, \mathbf{w} \in C_j} d(\mathbf{v}, \mathbf{w}). \tag{A.23}$$

De *volledige regel* schrijft voor dat:

$$f_k(C_i, C_j) = \max_{\mathbf{v} \in C_i, \mathbf{w} \in C_j} d(\mathbf{v}, \mathbf{w}). \tag{A.24}$$

Beide regels hebben hun voor- en nadelen. De *enkelvoudige regel* laat toe om niet-lineair scheidbare clusters te vormen, maar is bijzonder gevoelig aan ruis in de datacollectie. De *volledige regel* is minder gevoelig aan ruis, maar zal minder snel niet-lineaire clusters genereren.

A.6 Latente Dirichlet Allocatie

Latente Dirichlet Allocatie (LDA) is een generatief probabilistisch model voor het vormen van clusters met discrete data [131]. Omwille van het gebruik van LDA in Hoofdstuk 8 veronderstellen we hier tekstuele data. Een datacollectie is hier een verzameling van documenten en elk document wordt voorgesteld door een multiverzameling van woorden. Het aantal clusters wordt opnieuw als gekend verondersteld. Laat ons dit aantal opnieuw noteren als k.

LDA veronderstelt het bestaan van latente (d.z. verborgen) variabelen. Deze variabelen worden onderwerpen genoemd. Elk onderwerp komt overeen met één cluster, zodat er k onderwerpen worden verondersteld. Er wordt verder ook verondersteld dat elk document een mengsel van onderwerpen beschrijft. Als dusdanig kan er voor elk document een waarschijnlijkheidsverdeling worden geconstrueerd over het universum van onderwerpen. Een onderwerp wordt gemodelleerd als een waarschijnlijkheidsverdeling over het universum van woorden. De waarschijnlijkheidsverdeling over het universum van onderwerpen wordt voor het i^{de} document genoteerd als θ_i. De waarschijnlijkheidsverdeling over het universum van woorden wordt voor het j^{de} onderwerp genoteerd als φ_j. Voor zowel de verdelingen over het universum van onderwerpen als de verdelingen over het universum over woorden wordt de discrete Dirichlet verdeling vooropgesteld:

$$\Pr(\mathbf{x}|\alpha) = \frac{\Gamma\left(\sum_{i=1}^{n} \alpha_i\right)}{\prod_{i=1}^{n} \Gamma(\alpha_i)} \prod_{i=1}^{n} \left(\mathbf{x}_i^{\alpha_i - 1}\right) \tag{A.25}$$

met $\mathbf{x}$ een n-dimensionale willekeurige variabele, Γ de gamma functie en α een n-dimensionale parametervector met componenten strikt groter dan nul. Het schatten van de onbekende waarschijnlijkheden is een Bayesiaans inferentieprobleem. In Hoofdstuk 8 is gebruik gemaakt van *Gibbs sampling*. Dit is een iteratief proces waarbij in elke iteratie aan elk woord een onderwerp wordt toegekend. Onder de toekenning van woorden aan onderwerpen worden waarschijnlijksverdelingen θ_i en φ_j berekend voor documenten en onderwerpen.

Bijlage B

Leenvertalingen

Tabel B.1: Leenvertalingen

Nederlandse term	Engelse term
	bag
	bottom-up
	business intelligence
	data warehouse
	engine
	e-mail
	extract transform load process
	Gibbs sampling
	k-nearest-neighbour
	k-means
	match
	non-match
	overfitting
	prior
	record linkage
	top-down
affiene-gaten afstand	affine gap distance
boek van het leven	book of life
compleetheid	recall
databank	database
datamining	data mining
distributionele onzekerheidsvariabelen	distributional uncertainty variable

vervolgt op de volgende bladzijde

vervolg van de vorige bladzijde

Nederlandse term	Engelse term
enkelvoudige regel	single linkage rule
frequentistische	frequentistic
gemeenschapsgedreven	community driven
kardinaliteitsgebaseerd	cardinality based
lange staart	long tail
langste gemeenschappelijke deelsequentie	longest common subsequence
maximaal coherente deelverzamelingen	maximal coherent subsets
meervoudige documentsamenvatting	multiple document summarization
mogelijke verrassing	potential surprise
open bron	open source
overeenkomstpatroon	agreement pattern
schuivend venster	sliding window
slechtste-geval analyse	worst case analysis
standaardwaarde	default value
subjectivistisch	subjectivistic
superrelatie	superrelation
superverzameling	superset
tegenspraak	inconsistency
traag veranderende dimensies	slowly changing dimensions
traag veranderende eigenschappen	slowly changing properties
volledige regel	full linkage rule
wereldwijde web	world wide web
willekeurige variabele	random variable
XML opmaaktaal	XML markup language
zuiverheid	precision

Bijlage C

Afkortingen en acroniemen

Tabel C.1: Gebruikte afkortingen en acroniemen

BI	Business Intelligence
CR	Concept Relationeel
EM	Expectation Maximization
ETL	Extract Transform Load
LDA	Latente Dirichlet Allocatie
MDS	Meervoudige DocumentSamenvatting
NLP	Natural Language Processing
PCA	Principale Componenten Analyse
POI	Point Of Interest
RGB	Rood Groen Blauw
RSS	Really Simple Syndication
SIE	Smart Indexing Engine
TFIDF	Term Frequence Inverse Document Frequency
WWW	World Wide Web
XML	eXtendible Markup Language
bv.	bijvoorbeeld
d.i.	dit is
d.z.	dit zijn
resp.	respectievelijk

Bijlage D

Lijst met symbolen

D.1 Algemene symbolen

$\mathbb{R}$	Verzameling van de reële getallen
$\mathbb{N}$	Verzameling van de natuurlijke getallen
$\mathcal{P}$	Machtsverzameling

D.2 Boolese logica

$\mathbb{B}$	universum van Boolese waarheidswaarden
p	Boolese propositie
$\wedge$	Boolese conjunctie
$\vee$	Boolese disjunctie
$\neg$	Boolese negatie
$\Rightarrow$	Boolese implicatie

D.3 Vaagverzamelingen

$\mathcal{F}$	vage machtsverzameling
$\widetilde{V}$	vaagverzameling
$\mu_{\widetilde{V}}$	lidmaatschapsfunctie van $\widetilde{V}$
$\widetilde{V}_\alpha$	α-snede van $\widetilde{V}$
t	triangulaire norm
s	triangulaire conorm
co	complementoperator voor vaagverzamelingen
$\cap_t$	doorsnedeoperator voor vaagverzamelingen

vervolgt op de volgende bladzijde

vervolg van de vorige bladzijde	
$\cup_t$	unieoperator voor vaagverzamelingen
$f_{im,t}$	Residuele implicator
$f^{co}_{im,s}$	Residuele co-implicator

D.4 Multiverzamelingen

$\mathcal{M}$	multimachtsverzameling
$\mathcal{M}_n$	multimachtsverzameling voor kardinaliteit n
M	multiverzameling
ω_M	multipliciteitsfunctie van M
M_k	k-snede van M
$\langle M \rangle_k$	k-onderdrukking van M
$\cap$	doorsnedeoperator voor multiverzamelingen
$\cup$	unieoperator voor multiverzamelingen
$\oplus$	somoperator voor multiverzamelingen

D.5 Possibiliteitstheorie

Ω	uitkomstenverzameling
Π	possibiliteitsmaat
N	necessiteitsmaat
$\Pi(S\|C)$	conditionele possibiliteit
$N(S\|C)$	conditionele necessiteit
π	possibiliteitsverdeling
sur	verrassingsmaat
$\widetilde{p}$	possibilistische waarheidswaarde
$\widetilde{\neg}$	Zadeh-uitbreiding van $\neg$
$\widetilde{\wedge}$	Zadeh-uitbreiding van $\wedge$
$\widetilde{\vee}$	Zadeh-uitbreiding van $\vee$
$\widetilde{\Rightarrow}$	Zadeh-uitbreiding van $\Rightarrow$
$\widetilde{\wedge}_t$	t-uitbreiding van $\wedge$
$\widetilde{\vee}_t$	t-uitbreiding van $\vee$
$\widetilde{\Rightarrow}_t$	t-uitbreiding van $\Rightarrow$
Pos()	possibiliteit van een toekenning
Nec()	necessiteit van een toekenning
g_c	conjunctieve transformatie
g_d	disjunctieve transformatie
$\widetilde{\wedge}^*$	getransformeerde Zadeh-uitbreiding van $\wedge$
$\widetilde{\vee}^*$	getransformeerde Zadeh-uitbreiding van $\vee$
$\mathcal{A}$	actor die kennis over p postuleert

vervolgt op de volgende bladzijde

vervolg van de vorige bladzijde	
$\mathcal{A} \rightsquigarrow (p = T)$	actor $\mathcal{A}$ postuleert dat p waar is
$\mathcal{A} \rightsquigarrow (p = F)$	actor $\mathcal{A}$ postuleert dat p vals is
$\mathcal{A} = T$	verkorte notatie voor $\mathcal{A} \rightsquigarrow (p = T)$
$\mathcal{A} = F$	verkorte notatie voor $\mathcal{A} \rightsquigarrow (p = F)$
γ^T	vertrouwensmaat voor waar
γ^F	vertrouwensmaat voor vals
P_π	verzameling van possibilistische waarheidswaarden
$S_{\gamma^{T,F}}(P_\pi)$	Sugeno integraal voor P_π
$\cdot_{(i)^T}$	permutatie: ordening volgens zekerheid voor waar
$\cdot_{(i)^F}$	permutatie: ordening volgens zekerheid voor vals
T_S	Sugenotransformatie voor waar
F_S	Sugenotransformatie voor vals

D.6 Objecten en entiteiten

$\mathcal{E}$	universum van entiteiten
O	universum van complexe objecten
U	universum van atomaire objecten
$\mathcal{L}$	universum van labels
o	complex object
u	atomair object
ρ	referentiefunctie
lab	labelfunctie
λ	groeperingsfunctie
$\text{proj}_i(o)$	i^e deelobject van o
$\leftrightarrow$	coreferentierelatie
$\mathscr{M}$	meetproces voor objecten
$\mathscr{M}_i$	meetproces voor i^e eigenschap
$\perp_\forall$	beschrijving van een onbekende eigenschap
$\perp_\nexists$	beschrijving van een onbestaande eigenschap
ρ^*	veralgemeende referentiefunctie
$\leftrightarrow_\cap$	$\cap$-coreferentie
$\leftrightarrow_\subseteq$	$\subset$-coreferentie

D.7 Strings

$\mathcal{S}$	universum van strings
$\mathcal{S}_n$	universum van strings met lengte n
$\mathscr{A}$	alfabet
τ_s	karakteristieke functie van s
	vervolgt op de volgende bladzijde

vervolg van de vorige bladzijde	
I_n	indexverzameling $\{1, ..., n\}$
σ	lege string (lengte 0)
$\|s\|$	lengte van s
ind	indexfunctie voor strings
$s[\mathbf{q}]$	transformatie van s onder $\mathbf{q}$
$\sqsubset$	deelstringoperator
$\sqcap$	doorsnedeoperator voor strings
$\ominus$	verschiloperator voor strings
$\oplus$	concatenatieoperator voor strings
$\widehat{\sqsubset}$	zwakke deelstringoperator
$\widehat{\sqcap}$	zwakke doorsnedeoperator
$\widehat{\ominus}$	zwakke verschiloperator
len	karakteristieke lengte
$\mathscr{C}$	karakterverzameling
Δ	deelverschil tussen twee strings
$\mathscr{S}_{\mathscr{G}}$	splitsingsfunctie met grensverzameling $\mathscr{G}$
$\widehat{C}$	kandidaatverzameling

D.8 Documenten

$\mathcal{D}$	universum van documenten
$\mathcal{D}$	universum van concepten
$\psi(d)$	relationele transformatie van d
$\mathcal{T}$	transformatiefunctie
$\mathcal{C}(d)$	concepten van d
(c_1, c_2)	patroon: koppel van concepten
$(c_1, c_2) \in d$	(c_1, c_2) behoort tot d
$(c_1, c_2) \mathrel{\widehat{\in}} d$	(c_1, c_2) behoort bijna tot d
$(c_1, c_2) \mathrel{\widetilde{\in}} d$	(c_1, c_2) behoort mogelijks bijna tot d
$\mathcal{D}_{(c_1,c_2),\in}$	relationele selectie op basis van (c_1, c_2)
$\mathcal{D}_{(c_1,c_2),\,\widehat{\in}}$	conceptuele selectie op basis van (c_1, c_2)
$\mathcal{D}_{(c_1,c_2),\,\widetilde{\in}}$	$E_{\mathcal{C}}$-selectie op basis van (c_1, c_2)
$(c_1, c_2) \sim (c_1', c_2')$	afhankelijke patronen
$\mathrm{dep}((c_1, c_2), (c_1', c_2'))$	afhankelijkheidsgraad tussen patronen
$\mathbf{M}_{\mathbf{v}}$	afhankelijkheidsmatrix voor $\mathbf{v}$
$\mathbf{M}_{\mathbf{v}^*}$	afhankelijkheidsmatrix voor gepermuteerde $\mathbf{v}$
$\mathcal{E}_D$	entiteiten vertegenwoordigd in D
$\mathcal{V}$	verwerkingseenheid voor documenten
$\mathcal{B}(e)$	kost voor verwerking document dat e beschrijft
$H_{a,b}$	b^e veralgemeend harmonisch getal
Zipf	Zipf-verdeling

D.9 Evaluatoren en beslissingsmodellen

p_{o_1,o_2}	coreferentiële propositie van o_1 en o_2
E_O	evaluator over O
$E_{O,R}$	semantische evaluator over O
sel	selectie van koppels
ι	één-op-één afbeelding tussen verzamelingen
$\mathbf{M}_{A,B}$	matrix met possibilistische waarheidswaarden
$<_{\text{leximax}}$	leximax orderelatie voor $\mathcal{F}(\mathbb{B})^n$
$q_{\alpha,\beta,\delta}$	geparameteriseerde kwantorfunctie
$E_{\mathcal{S}}$	één-niveau evaluator voor strings
$E^*_{\mathcal{S}}$	twee-niveau evaluator voor strings
$E_{\mathcal{D}}$	evaluator voor documenten
$E^{\text{op}}_{\mathcal{D}}$	op-evaluator voor documenten
$\mathscr{B}$	tweewaardig beslissingsmodel
$\mathscr{D}$	driewaardig beslissingsmodel
R^+	transitieve sluiting van R
R^-	transitieve opening van R
$R^T_{\mathscr{B}}$	coreferente koppels volgens $\mathscr{B}$
$R^F_{\mathscr{B}}$	niet-coreferente koppels volgens $\mathscr{B}$
$R^T_{\mathscr{D}}$	coreferente koppels volgens $\mathscr{D}$
$R^F_{\mathscr{D}}$	niet-coreferente koppels volgens $\mathscr{D}$
$R^{T,F}_{\mathscr{D}}$	onzekere koppels volgens $\mathscr{D}$
$\mathscr{P}_{R_{\mathscr{B}}}$	partitie van objecten op basis van $R_{\mathscr{B}}$
$\mathscr{P}^*_{R_{\mathscr{B}}}$	minimale partitie van objecten op basis van $R_{\mathscr{B}}$
$\mathcal{H}_{f_{S_{\gamma^{T,F}}}}$	Sugeno-gebaseerde partitiegenerator
$[\![O]\!]$	maximale en voldoende klasse van O
$\mathscr{H}$	maximale en voldoende partitiegenerator

D.10 Samenvoeging

ϖ_O	samenvoegingsfunctie over O
$\mathbf{r}^{\tilde{n}}$	rechtsprojectie van $\tilde{n}$
$\mathbf{l}^{\tilde{n}}$	linksprojectie van $\tilde{n}$
$\mathbf{l}^{\tilde{n},\text{rev}}$	gereverteerde linksprojectie van $\tilde{n}$
$\prec_{\text{sup}}$	sup-orderelatie voor $\mathcal{F}(\mathbb{N})$
$\prec_{\text{inf}}$	inf-orderelatie voor $\mathcal{F}(\mathbb{N})$
$\varpi_{O,R}$	semantische samenvoegingsfunctie over O

Bijlage E

Datacollecties

In deze bijlage worden ter illustratie korte steekproeven van de gebruikte datacollecties gegeven. Voor de datacollecties gebruikt in Hoofdstuk 6 en Hoofdstuk 7 maken we onderscheid tussen datacollecties afkomstig uit één bron en datacollecties afkomstig uit twee bronnen. We vermelden voor elk object steeds een 'eid'. Objecten met een gelijke waarde voor 'eid' zijn coreferent. Objecten met een verschillende waarde voor 'eid', zijn niet coreferent.

E.1 Datacollecties gebruikt in Hoofdstuk 6

E.1.1 Datacollectie 'people'

Datacollectie 'people' is een datacollectie bestaande uit twee bronnen. Beide bronnen beschrijven personen aan de hand van hun naam.

Tabel E.1: People: bron 1

s	eid
abbott, yvonne	1
abughrien, badia	2
acton, ciara	3
ms. gillian acton, secretary	4
adams, adrian	5

Tabel E.2: People: bron 2

s	eid
mrs. yvonne abbott	1
ms. badia abughrien	2
vervolgt op de volgende bladzijde	

vervolg van de vorige bladzijde	
ciara acton	3
acton, gillian	4
dr. suren aghajanian	6

E.1.2 Datacollectie 'bird1'

Datacollectie 'bird1' is een datacollectie bestaande uit twee bronnen. Beide bronnen beschrijven vogels aan de hand van hun soortnaam.

Tabel E.3: Bird1: bron 1

s	eid
baltimore oriole (icterus galbula)	1
red-eyed vireo (vireo olivaceus)	2
carolina wren(thryothorus ludovicianus)	3
american kestrel (falco sparverius)	4
black vulture (coragyps atratus)	5

Tabel E.4: Bird1: bron 2

s	eid
baltimore oriole (icterus parisorum)	1
(vireo olivaceus)	2
carolina wren	3
brown creeper	6
ovenbird	7

E.1.3 Datacollectie 'bird2'

Datacollectie 'bird2' is een datacollectie bestaande uit twee bronnen. Beide bronnen beschrijven vogels aan de hand van hun soortnaam.

Tabel E.5: Bird2: bron 1

s	eid
tufted titmouse	1
canada goose	2
broad-winged hawk	3
northern cardinal	4
american goldfinch	5

Tabel E.6: Bird2: bron 2

s	eid
tufted titmouse	1
canada goose	2
broad-winged hawk	3
cooper's hawk	6
northern goshawk	7

E.1.4 Datacollectie 'bird3'

Datacollectie 'bird3' is een datacollectie bestaande uit twee bronnen. Beide bronnen beschrijven vogels aan de hand van hun soortnaam.

Tabel E.7: Bird3: bron 1

s	eid
bittern, american (butor d'amérique): botaurus lentiginosus	1
avocet, american (avocette d'amerique): recurvirostra americana	2
owl: barn	3
dove: inca	4
booby, masked	5

Tabel E.8: Bird3: bron 2

s	eid
american bittern	1
american avocet	2
barn owl	3
black-bellied whistling-duck	6
black-necked stilt	7

E.1.5 Datacollectie 'bird4'

Datacollectie 'bird4' is een datacollectie bestaande uit twee bronnen. Beide bronnen beschrijven vogels aan de hand van hun soortnaam.

Tabel E.9: Bird4: bron 1

s	eid
ibis: white-faced plegadic chihi	1
stork, wood: mycteria americana	2
bittern, american (butor d'amérique): botaurus lentiginosus	3
egret: great (grande aigrette) casmerodius albus	4

vervolgt op de volgende bladzijde

vervolg van de vorige bladzijde	
egret: snowy (aigrette niegeuse) egretta thula	5

Tabel E.10: Bird4: bron 2

s	eid
white ibis eudocimus albus	1
wood stork mycteria americana	2
american bittern botaurus lentiginosus	3
cattle egret bubulcus ibis	6
sora porzana carolina	7

E.1.6 Datacollectie 'game'

Datacollectie 'game' is een datacollectie bestaande uit twee bronnen. Beide bronnen beschrijven computerspellen aan de hand van hun naam.

Tabel E.11: Game: bron 1

s	eid
3-d movie maker	1
arthur's birthday	2
catz	3
chill manor	4
creative writer	5

Tabel E.12: Game: bron 2

s	eid
microsoft 3d moviemaker	1
arthur's birthday	2
hollywood high	6
how to draw cartoons	7
how to draw the marvel way	8

E.1.7 Datacollectie 'park'

Datacollectie 'park' is een datacollectie bestaande uit twee bronnen. Beide bronnen beschrijven parken aan de hand van hun naam.

Tabel E.13: Park: bron 1

s	eid
vervolgt op de volgende bladzijde	

vervolg van de vorige bladzijde	
acadia np	1
agate fossil beds nm	2
allegheny portage railroad nhs	3
american memorial park	4
amistad nra	5

Tabel E.14: Park: bron 2

s	eid
acadia np	1
agate fossil beds nm	2
adams nhs	6
alagnak wild river	7
alibates flint quarries nm	8

E.1.8 Datacollectie 'animal'

Datacollectie 'animal' is een datacollectie bestaande uit twee bronnen. Beide bronnen beschrijven diersoorten aan de hand van hun soortnaam.

Tabel E.15: Park: bron 1

s	eid
salamander, cheat mountain	1
salamander, desert slender	2
toad, houston	3
sensitive joint-vetch	4
sensitive joint-vetch	4

Tabel E.16: Park: bron 2

s	eid
cheat mountain salamander	1
desert slender salamander	2
houston toad	3
golden-crowned sparrow	5
white-crowned sparrow	6

E.1.9 Datacollectie 'census'

Datacollectie 'census' is een datacollectie bestaande uit twee bronnen. Beide bronnen beschrijven personen aan de hand van hun naam en adres.

Tabel E.17: Census: bron 1

s	eid
bell glorianne, m 7111 3rd	1
scchwartz john, f 7121 3rd	2
scchwartz tricia, e 7121 3rd	3
hubric charles, p 7127 3rd	4
hubric diane, 7127 3rd	5

Tabel E.18: Census: bron 2

s	eid
schwartz patricia e 7121 3rd	3
hubrick charls p 7127 3rd	4
pollock andrew 7139 3rd	5
pollock mirian j 7139 3rd	6
pollock jamal f 7139 3rd	7

E.1.10 Datacollectie 'univ'

Datacollectie 'univ' is een datacollectie bestaande uit één bron. De bron beschrijft universiteiten aan de hand van hun naam.

Tabel E.19: Univ

s	eid
science applications, inc., mclean	1
science applications international corporation, mclean	1
analytical mechanics associates, inc., hampton	2
west virginia university, morgantown	3
west virginia university, morgantown, w.	3

E.1.11 Datacollectie 'streets5'

Datacollectie 'streets5' is een datacollectie bestaande uit twee bronnen. Beide bronnen beschrijven straten aan de hand van hun naam.

Tabel E.20: Streets5: bron 1

s	eid
buerstedelei	1
kamersveld	2
asiadok-noordkaai	3
baron dhanislaan	4

vervolgt op de volgende bladzijde

vervolg van de vorige bladzijde	
beeckmansstraat	5

Tabel E.21: Streets5: bron 2

s	eid
bueerstedelei	1
kemersveld	2
asiadok-noorrdkaai	3
baron dhaislaan	4
beckmansstraat	5

E.1.12 Datacollectie 'constraint'

Datacollectie 'constraint' is een datacollectie bestaande uit één bron. De bron beschrijft wetenschappelijke publicaties aan de hand van onder andere hun titel en hun auteurs.

Tabel E.22: Constraint

s	eid
antonisse, j (1989). a new interpretation of schema notation that overturns the binary encoding constraint. icga iii. san mateo, ca, morgan kaufmann.	1
antonisse, j., a new interpretation of schema encoding that overturns the binary encoding constraint	1
minton, s. 1993. integrating heuristics for constraint satisfaction problems: a case study. in proceedings of the eleventh national conference on artificial intelligence, 120-126.	2
s. minton. integrating heuristics for constraint satisfaction problems: a case study. in aaai proceedings, 1993.	2
minton, s. (1984). constraint-based generalization - learning game-playing plans from single examples. in proceedings of the fourth national conference on artificial intelligence. william kaufmann, 251-254. mitchell, t., allen, j., chalasani, p., cheng,	3

E.1.13 Datacollectie 'face'

Datacollectie 'face' is een datacollectie bestaande uit één bron. De bron beschrijft wetenschappelijke publicaties aan de hand van onder andere hun titel en hun auteurs.

Tabel E.23: Face

vervolgt op de volgende bladzijde

vervolg van de vorige bladzijde	
s	eid
laurel, b. (1990) interface agents: metaphors with character, in the art of humancomputer interface design, ed. b. laurel, addison wesley: reading, 355-365.	1
laurel, b., interface agents: metaphors with character, in the art of human computer interface design, b. laurel (ed), addisonwesley, 1990.	1
donald p. mckay, timothy w. finin, and anthony o'hare. the intelligent database interface: integrating ai and database systems. in aaai-90: proceedings of the eighth national conference on artificial intelligence, 1990.	2
maes, p., and kozierok, r. 1993. learning interface agents. in proceedings of aaai-93.	3
maes, p., kozierok, r. (1993). learning interface agents. in: proceedings of the aaai'93.	3

E.1.14 Datacollectie 'reasoning'

Datacollectie 'reasoning' is een datacollectie bestaande uit één bron. De bron beschrijft wetenschappelijke publicaties aan de hand van onder andere hun titel en hun auteurs.

Tabel E.24: Reasoning

s	eid
p. langley and s. sage. oblivious decision trees and abstract cases. in proceedings of the 1994 aaai workshop on casebased reasoning, pages 113-117, seattle, ca, 1994. aaai press.	1
langley, p. and sage, s. (1994b), oblivious decision trees and abstract cases, in working notes of the aaai-94 workshop on case-based reasoning, aaai press, seattle, pp. 113-117.	1
langley, p., and sage, s. 1994. oblivious decision trees and abstract cases. in working notes of the aaai94 workshop on case-based reasoning. in press.	1
p. langley, s. sage, oblivious decision trees and abstract cases, to be presented at the aaai94workshop on case-based reasoning, seattle, wa., 1994.	2
d. w. aha and r. l. bankert. feature selection for case-based classification of cloud types: an empirical comparison. in proceedings of the 1994 aaai workshop on casebased reasoning, pages 106-112, seattle, wa, 1994. aaai press.	3

E.1.15 Datacollectie 'reinforcement'

Datacollectie 'reinforcement' is een datacollectie bestaande uit één bron. De bron beschrijft wetenschappelijke publicaties aan de hand van onder andere hun titel en hun auteurs.

Tabel E.25: Reinforcement

s	eid
singh, s., scaling reinforcement learning algorithms by learning variable temporal resolution models. in proc. of the ninth international workshop on machine learning, 1992.	1
singh, s., scaling reinforcement learning algorithms by learning variable temporal resolution models. in proc. of the ninth international workshop on machine learning, 1992.	1
r. andrew mccallum. using transitional proximity for faster reinforcement learning. in the proceedings of the ninth international machine learning conference. morgan kaufmann publishers, inc., 1992.	2
jeffery a. clouse and paul e. utgoff. a teaching method for reinforcement learning. in the proceedings of the ninth international machine learning conference. morgan kaufmann publishers, inc., 1992.	3
clouse, j. and utgoff, p. (1992). a teaching method for reinforcement learning. in machine learning: proceedings of the ninth international workshop, (pp. 92-101), aberdeen, scotland.	3

E.1.16 Datacollectie 'restaurant'

Datacollectie 'restaurant' is een datacollectie bestaande uit twee bronnen. De beide bronnen beschrijven restaurants aan de hand van hun naam.

Tabel E.26: Restaurant: bron 1

s	eid
art's delicatessen	1
restaurant katsu	2
postrio	3
l'orangerie	4
ritz-carlton restaurant and dining room	5

Tabel E.27: Restaurant: bron 2

s	eid
vervolgt op de volgende bladzijde	

vervolg van de vorige bladzijde	
art's deli	1
katsu	2
fleur de lys	6
fringale	7
hawthorne lane	8

E.2 Datacollecties gebruikt in Hoofdstuk 7

E.2.1 Datacollectie 'census'

Datacollectie 'census' is een datacollectie bestaande uit twee bronnen. Beide bronnen beschrijven personen aan de hand van hun voornaam, middennaam, naam, straat en huisnummer.

Tabel E.28: Census: bron 1

voornaam	middennaam	naam	straat	huisnummer	eid
Glorianne	M	Bell	3rd	7111	1
John	F	Scchwartz	3rd	7121	2
Tricia	E	Scchwartz	3rd	7121	3
Charles	P	Hubric	3rd	7127	4
Diane	$\perp$	Hubric	3rd	7127	5

Tabel E.29: Census: bron 2

voornaam	middennaam	naam	straat	huisnummer	eid
Patricia	E	Schwartz	3rd	7121	3
Charls	P	Hubrick	3rd	7127	4
Andrew	$\perp$	Pollock	3rd	7139	5
Mirian	J	Pollock	3rd	7139	6
Jamal	F	Pollock	3rd	7139	7

E.2.2 Datacollectie 'cora'

Datacollectie 'cora' is een datacollectie bestaande uit één bron. Deze bron beschrijft publicaties aan de hand van hun titel, auteurs, pagina's, bron en jaar.

Tabel E.30: Cora

titel	auteurs	pagina's	bron	jaar	eid
vervolgt op de volgende bladzijde					

vervolg van de vorige bladzijde					
on-line prediction and conversion strategies.	n. cesa-bianchi, y. freund, d. p. helmbold, and m. warmuth.	205-216	in computational learning theory	⊥	1
on-line prediction and conversion strategies.	n. cesa-bianchi, y. freund, d.p. helmbold, and m. warmuth.	205-216	in eurocolt	1993	1
learning to order things	william w. cohen, robert e. schapire, and yoram singer	⊥	in advances in neural information processing systems 10	1998	2
learning to order things	w. w. cohen, r. e. shapire, and y. singer.	⊥	in advances in neural information processing systems 10	1998	2
on the complexity of teaching	s. goldman and m. kearns	303-314	in colt-91	1991	3

E.2.3 Datacollectie 'hotels'

Datacollectie 'hotels' is een datacollectie bestaande uit twee bronnen. Beide bronnen beschrijven hotels aan de hand van hun naam, adres, sterren en stad.

Tabel E.31: Hotels: bron 1

naam	adres	sterren	stad	eid
radisson blu park lane hotel	van eycklei 34, 2018 antwerpen	5	Antwerpen	1
hotel helios	zeedijk 92, 8370 blankenberge	4	Blankenberge	2
hotel richmond thonnon	j. van maerlantstraat 79, 8370 blankenberge	4	Blankenberge	3
beach palace hotel	zeedijk 77-79, 8370 blankenberge	4	Blankenberge	4
hotel aazaert ****	hoogstraat 25, 8370 blankenberge	4	Blankenberge	5

Tabel E.32: Hotels: bron 2

naam	adres	sterren	stad	eid
radisson blu park lane hotel	van eycklei 34, antwerpen	5	Antwerpen	1
golden tulip antwerp centre	lange kievitstraat 125, antwerpen	4	Antwerpen	6
helios	zeedijk, 92, blankenberge	4	Blankberge	2
best western hotel richmond thonnon	van maerlantstraat, 79, blankenberge	4	Blankberge	3
beach palace	zeedijk 77 - 79, blankenberge	4	Blankberge	4

E.2.4 Datacollectie 'restaurant'

Datacollectie 'restaurant' is een datacollectie bestaande uit twee bronnen. De beide bronnen beschrijven restaurants aan de hand van hun naam, straat, stad en type.

Tabel E.33: Restaurant: bron 1

naam	straat	stad	type	eid
art's delicatessen	12224 ventura blvd.	studio city	american	1
restaurant katsu	1972 n. hillhurst ave.	los angeles	asian	2
postrio	545 post st.	san francisco	american	3
l'orangerie	903 n. la cienega blvd.	los angeles	french	4
ritz-carlton restaurant and dining room	600 stockton st.	san francisco	american	5

Tabel E.34: Restaurant: bron 2

naam	straat	stad	type	eid
art's deli	12224 ventura blvd.	studio city	delis	1
katsu	1972 hillhurst ave.	los feliz	japanese	2
fleur de lys	777 sutter st.	san francisco	french	6
fringale	570 fourth st.	san francisco	french	7
hawthorne lane	22 hawthorne st.	san francisco	californian	8

E.3 Datacollecties gebruikt in Hoofdstuk 8

In Hoofdstuk 8 zijn experimenten gedaan met betrekking tot het clusteren van coreferente teksten. We geven hier vijf teksten uit deze datacollectie, telkens voorafgegaan door hun onderwerp. Voor de opbouw van de datacollectie zijn volgende websites geraadpleegd: www.standaard.be, www.deredactie.be, www.tijd.be, www.nd.nl, www.parool.nl, www.nu.nl en www.telegraaf.nl.

Tekst 1 (Crash Schiphol)

"Toestel Turkish Airlines had eerder problemen.

ANKARA - De neergestorte Boeing 737-800 van Turkish Airlines had enkele dagen voor het ongeluk bij de Polderbaan op Schiphol tot tweemaal toe technische problemen. Dat meldde het Turkse persbureau Dogan donderdag. Het eerste mankement werd opgemerkt tijdens een vlucht op 18 februari. Volgens de piloten waren er problemen met de vleugels van het toestel. De Boeing heeft daarna een technische controle ondergaan in de hangar van Turkish Airlines op het Ataturk-vliegveld in Istanbul. De tweede storing had twee dagen voor het ongeluk plaats. De piloten moesten toen de start afbreken. De Turkse minister van Transport Binali Yildirim meldde woensdag dat de Boeing 737-800 geen technische mankementen had. Dat zou blijken uit de laatste technische controle door de Turkse burgerluchtvaartorganisatie op 22 december vorig jaar."

Tekst 2 (Crash Schiphol)

"Wat gebeurde er in de minuten voor de crash?.

do 26/02/09 11:00 (UPDATE audio) - Het blijft gissen naar de oorzaak van de vliegtuigcrash van gisteren in de buurt van Schiphol. Uit de communicatie tussen de cockpit en de verkeerstoren blijkt dat er geen radiocontact is geweest in de minuten voor de crash. Op de website www.liveatc.net staan de gesprekken tussen de controletoren en de vliegtuigen in de minuten voor en na de crash. Opvallend is dat er geen contact was met het vliegtuig van Turkish Airlines zelf. Andere piloten hadden de verkeerstoren wel gewaarschuwd dat de Boeing 737 vreemde vliegbewegingen maakte. Het onderzoek van de zwarte dozen moet nu meer duidelijkheid verschaffen. Daaruit zal onder meer blijken welke manoeuvres het toestel nog heeft uitgevoerd voor de crash en op welke plaatsen de piloten zaten. De piloten zouden er wel alles aan gedaan hebben om de landingsbaan te bereiken en hebben pas op het einde besloten om een noodlanding te maken in een veld. Intussen doen allerlei geruchten de ronde. Een getuige zag dat het vliegtuig net voor de crash nog even steeg en dan plots naar beneden stortte. Volgens experts kan dat betekenen dat de motoren plots zijn uitgevallen omdat er niet genoeg kerosine aan boord was."

Tekst 3 (Crash Buffalo)

"Doden door vliegtuigongeluk bij Buffalo.

WASHINGTON - Een tweemotorig passagiersvliegtuig met 48 inzittenden aan boord is donderdag neergestort op een huis in de buurt van het vliegveld van Buffalo in de Amerikaanse staat New York. Alle 44 passagiers en vier bemanningsleden kwamen om. Zeker een persoon op de grond liet het leven. Dat heeft de politie bevestigd. Persbureau AP toont foto's op haar website, CNN heeft videobeelden van de brand."

Tekst 4 (Dopingzaak Valverde)

"Valverde krijgt hulp van Spaanse rechter.

ROME/MADRID - Het Italiaans olympisch comité CONI mag geen gebruikmaken van uit Madrid verkregen bewijsmateriaal, waaronder bloedstalen, tegen de Spaanse wielrenner Alejandro Valverde. Dat heeft de Spaanse onderzoeksrechter Antonio Serrano, die belast is met de zaak 'Operacion Puerto', woensdag verklaard. Valverde moet zich in Itali verantwoorden voor mogelijke betrokkenheid bij de dopingaffaire. Het verhoor bij het CONI is verplaatst naar donderdag. Eerder had Valverde om uitstel gevraagd naar vrijdag om zijn advocaat in de gelegenheid te stellen zich terdege voor te bereiden.

Hij krijgt nu hulp van de onderzoeksrechter in Madrid, die stelt dat de bewijsmaterialen niet ter beschikking van het CONI hadden mogen worden gesteld. Serrano leidt het heropende onderzoek naar het dopingschandaal rond de arts Eufemiano Fuentes, waabij tal van wielrenners maar ook veel andere sporters betrokken zouden zijn. Het CONI zou over bewijs beschikken dat Valverde een van de klanten van Fuentes is geweest.

De wielrenner moet zich donderdag in Rome komen verantwoorden. Hij liet na de interventie van Serrano open of hij daadwerkelijk in Itali verschijnt. 'Valverde zal met zijn advocaten overleggen en dat doen wat zij hem aanraden', verklaarde zijn manager Paco Snchez.

Volgens de Spaanse onderzoeksrechter mogen de Italiaanse dopingjagers de bloedstalen uit het laboratorium van Fuentes niet gebruiken, omdat die dienen als bewijsmateriaal in de zaak tegen de dopingarts. Zo lang daarin geen uitspraak is gedaan, kan het materiaal niet in een andere zaak worden gebruikt. Daarbij stelt Serrano dat het CONI geen juridische instelling is.

Van oud-ploeggenoot Jesus Manzano hoeft Valverde het in ieder geval niet te hebben. Manzano bevestigde woensdag in de Spaanse sportkrant AS dat dopingpraktijken bij de ploeg Kelme in 2002 en 2003 gemeengoed waren. 'Ze dienden Valverde dezelfde middelen toe als mij', verklaarde Manzano, die zei inmiddels door het CONI te zijn verhoord. 'In Itali pakken ze de zaken serieus aan. Als Ivan Basso bestraft moet worden, doen ze dat gewoon. Als er in Spanje een grote naam bij betrokken is, wordt de zaak gewoon gesloten.' "

Tekst 5 (Dopingzaak Valverde)

"Bloedstaal is van Valverde.

ROME - Een van de bloedstalen die zijn gevonden in het laboratorium van de Spaanse dopingarts Eufemiano Fuentes is van Alejandro Valverde. Dat zei Ettore Torri, de antidopingprocureur van het CONI, donderdag tijdens het verhoor van de Spaanse wielrenner in Rome. We kunnen met zekerheid zeggen dat het bloed in zak nummer achttien van Valverde afkomstig is, aldus Torri.

Tijdens de Ronde van Frankrijk van 2008 werden bij Valverde een bloed- en urinemonster afgenomen. De renner testte niet positief, maar uit dna-onderzoek zou zijn gebleken dat het monster van dezelfde persoon afkomstig was als een bloedstaal die werd aangetroffen in de praktijk van de omstreden Fuentes, de spil in het dopingschandaal Operacion Puerto. 'We beschikken ook over documenten die naar Valverde refereren en hebben bewijzen dat hij geld-

bedragen aan Fuentes overgemaakt heeft', zei Torri verder. 'Zijn advocaten hebben twee weken de tijd om hun verdediging voor te bereiden.'

Valverde is in Rome in gezelschap van twee advocaten, zijn ploegleider Eusebio Unzue en zijn manager Antonio Sanchez. 'Alejandro is onschuldig', verklaarde zijn advocaat Federico Cecconi meteen na de zitting, waar Valverde zelf weinig spraakzaam was. 'Dat hij ook voor het gerecht moet verschijnen, beschouwt hij als een verplichte procedure. We zijn hier naartoe gekomen om aan te tonen dat Valverde hier niets mee van doen heeft.'

Eerder op de dag kondigde de openbaar aanklager in Rome een onderzoek aan naar vermeende dopingpraktijken van Valverde. Ook loopt er een procedure bij het internationaal sporttribunaal in Lausanne (CAS). Die zaak is aangespannen door het mondiale antidopingbureau WADA. Dat bureau vindt dat justitie in Spanje Operacion Puerto in de doofpot probeert te stoppen en wil dat via het CAS voorkomen."

Bibliografie

[1] Edgar Frank Codd. *A Relational Model of Data for Large Shared Data Banks*. Communications of the ACM, 13(6):377–387, 1970.

[2] Alonzo Church. *A set of postulates for the foundation of logic*. Annals of Mathematics, 2(33):346–366, 1932.

[3] Georg Cantor. *Uber eine Eigenschaft des Inbegriffes aller reellen algebraischen Zahlen*. Crelles Journal, pages 258 – 262, 1874.

[4] Abraham Fraenkel. *Axiomatische Theorie der geordneten Mengen*. Journal of Pure and Applied Mathematics, pages 129–158, 1926.

[5] Ernst Zermelo. *Grenzzahlen und Mengenbereiche*. Fundamenta Mathematicae, pages 29–47, 1930.

[6] Lotfi Zadeh. *Fuzzy sets*. Information and Control, 8:338–353, 1965.

[7] Didier Dubois and Henri Prade. *Fundamentals of Fuzzy Sets*. Kluwer Academic, 2000.

[8] Gert De Cooman. *Towards a possibilistic logic. In: Fuzzy Set Theory and Advanced Mathematical Applications, edited by Da Ruan, Kluwer Academic, pp. 89–133, Boston.*, 1995.

[9] Ronald Yager. *On the theory of bags*. International Journal of General Systems, 13(1):23–27, 1986.

[10] Donald Knuth. *The Art of computer programming Volume 2 (Seminumerical Algorithms)*. Addison-Wesley, 1981.

[11] T. Yohanna D. Singh, A Ibrahim and J. N. Singh. *An overview of the applications of multisets*. Novi Sad Journal of Mathematics, 37(2):73–92, 2007.

[12] Wayne Blizard. *Generalizations of the Concept of Set: A Formal Theory of Multisets*. PhD thesis, Mathematical Institute, University of Oxford, 1986.

[13] Wayne D. Blizard. *Multiset Theory.* Notre Dame Journal of Formal Logic, 30(1), 1989.

[14] Wayne D. Blizard. *The Development of Multiset Theory.* Modern Logic, 1:319–352, 1991.

[15] Wayne D. Blizard. *Dedekind multisets and function shells.* Theoretical Computer Science, 110:79–98, 1993.

[16] Lotfi Zadeh. *Fuzzy sets as a basis for a theory of possibility.* Fuzzy Sets and Systems, 1:3–28, 1978.

[17] Brian Gaines and Ladislav Kohout. *Possible automata.* In Proceedings of the International symposium on multiple-valued logic, pages 183–196, Bloomington, USA, 1975.

[18] Didier Dubois and Henri Prade. *Possibility Theory.* Plenum Press, 1988.

[19] Glenn Shafer. *A mathematical theory of evidence.* Princeton University Press, 1961.

[20] Peter Walley. *Statistical Reasoning with Imprecise Probabilities.* Chapman and Hall, 1991.

[21] George Shackle. *Decision, order and time in human affairs.* Cambridge University Press, 1961.

[22] Gert De Cooman. *Evaluatieverzamelingen en - afbeeldingen. Een ordetheoretische benadering van vaagheid en onzekerheid.* PhD thesis, Faculteit Toegepaste Wetenschappen, Universiteit Gent, 1993.

[23] Ellen Hisdal. *Conditional possibilities independence and noninteraction.* Fuzzy Sets and Systems, 1:283–297, 1978.

[24] Didier Dubois and Henri Prade. chapter Formal representations of uncertainty, pages 85–156. Wiley, 2009.

[25] Stephen Kleene. *On notation for ordinal numbers.* Journal of symbolic logic, 3(4):150–155, 1938.

[26] Stephen Kleene. *Introduction to metamathematics.* Van Nostrand, 1952.

[27] Nicolas Rescher. *Many-valued logic.* McGraw-Hill, 1969.

[28] Henri Prade. *Possibility sets, fuzzy sets and their relation to Lukasiewicz logic.* In Proceedings of the International Symposium on Multiple-Valued Logic, pages 223–227, 1982.

[29] Antwan Van Schooten. *Ontwerp en implementatie van een model voor de representatie en manipulatie van onzekerheid en imprecisie in databanken en expert systemen.* PhD thesis, Ghent University, 1988.

[30] Amihai Motro and Igor Rakov. *Estimating the Quality of Data in Relational Databases.* In Proceedings of the International Conference on Information Quality, pages 94–106, 1996.

[31] Roderick Cattel, Douglas Barry, Mark Berler, Jeff Eastman, David Jordan, Craig Russel, Olaf Schadow, Torsten Stanienda, and Fernando Velez, editors. *The Object Data Standard: ODMG 3.0.* Morgan Kaufmann, 2000.

[32] Tim Bray, Jean Paoli, C. M. Sperberg-McQueen, Eve Maler, and Franois Yergeau. *Extensible Markup Language (XML) 1.0*, 2008.

[33] RamezElmasri and Shamkant Navathe. *Fundamentals of database systems.* Pearson, 2007.

[34] Erhard Rahm and Philip Bernstein. *A survey of approaches to automatic schema matching.* The VLDB journal, 10:334–350, 2001.

[35] James Larson, Shamkant Navathe, and Ramez Elmasri. *A Theory of Attribute Equivalence in Databases with Application to Schema Integration.* IEEE Transactions on Software Engeneering, 15(4):449–463, 1989.

[36] Vincenç Torra and Jordi Nin. *Record linkage for database integration using fuzzy integrals.* International Journal of Intelligent Systems, 23:715–734, 2008.

[37] Tatiana Jaworska. *Object extraction as a basic process for content-based image retrieval (CBIR) system.* Opto-Electronics Review, 15(4):184–195, 2007.

[38] Ralph Kimball and Margy ross. *The Data Warehouse Toolkit.* Wiley, 2002.

[39] Henri Prade and Claudette Testemale. *Generalizing Database Relational Algebra for the Treatment of Incomplete/Uncertain Information and Vague Queries.* Information Science, 34(2):115–143, 1984.

[40] Amihai Motro. *Management of Uncerainty in database Systems.* In Modern Database Systems, pages 457–476. 1995.

[41] Amihai Motro. *Sources of Uncertainty, Imprecision, and Inconsistency in Information Systems.* In Uncertainty Management in Information Systems, pages 9–34. 1996.

[42] Ivan Fellegi and Alan Sunter. *A Theory for Record Linkage.* American Statistical Association Journal, 64(328):1183–1210, 1969.

[43] *Interim Report: ANSI/X3/SPARC Study Group on Data Base Management Systems.* Bulletin of ACM SIGMOD, 7(2):1–140, 1975.

[44] Edgar Frank Codd. *Understanding relations.* Bulletin of ACM SIGMOD, 7(3-4):23–28, 1975.

[45] Edgar Frank Codd. *RM/T: Extending the relational model to capture more meaning.* ACM Transactions on Database Systems, 4(4):397–434, 1979.

[46] Edgar Frank Codd. *Missing information (applicable and inapplicable) in relational databases.* ACM SIGMOD Record, 15(4):53–78, 1986.

[47] Guy De Tré, Rita De Caluwé, and Henri Prade. *Null values in fuzzy databases.* Journal of Intelligent Information Systems, 30:93 – 114, 2008.

[48] Ross Quillian. *Semantic Information Processing*, chapter Semantic memory, pages 227–270. MIT Press, 1968.

[49] Allan Collins and Ross Quillian. *Retrieval time from semantic memory.* Journal of Verbal Learning and Verbal Behavior, 8:240–247, 1969.

[50] Allan Collins and Ross Quillian. *Organization of memory*, chapter How to make a language user., pages 309–351. Academic Press, 1972.

[51] J Raaijmakers and M Schiffrin. *Search of associative memory.* Psychological Review, 8(2):98–134, 1981.

[52] J Anderson. *The Architecture of Cognition.* Harvard University Press, 1983.

[53] T Landauer and S Dumais. *A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge.* Psychological Review, 104:211–240, 1997.

[54] K Lund, C Burgess, and R Atchley. *Semantic and associative priming in a high-dimensional semantic space.* In Cognitive Science Proceedings (LEA), pages 660–665, 1995.

[55] K Lund and C Burgess. *Producing high-dimensional semantic spaces from lexical co-occurrence.* Behavior Research Methods, Instruments and Computers, 28(2):203–208, 1996.

[56] Roy Rada, Hafedh Mili, Ellen Bicknell, and Maria Blettner. *Developments and applications of a metric on semantic nets.* IEEE Transactions on Systems, Man and Cybernetics, 19(1):17–30, 1989.

[57] Philip Resnik. *Using Information Content to Evaluate Semantic Similarity in a Taxonomy.* In Proceedings of the 14th International Joint Conference on Artificial Intelligence, pages 448–453, 1995.

[58] Jay Jiang and David Conrath. *Semantical similarity based on corpus statistics and lexical taxonomy.* In Proceedings of International Conference Research on Computational Linguistics, pages 15–33, 1997.

[59] Yuhua Li. *An approach for measuring Semantic Similarity between Words using multiple information sources.* IEEE Transactions on Knowledge and Data Engineering, 15(4):871–882, 2003.

[60] Guy De Tré and Bernard de Baets. *Aggregating Constraint Satisfaction Degrees Expressed by Possibilistic Truth Values.* IEEE Transactions of Fuzzy Systems, 11(3):361–368, 2003.

[61] Michio Sugeno. *Theory of fuzzy integrals and it's applications.* PhD thesis, Tokyo Institute of Technology, 1974.

[62] Lotfi Zadeh. *A computational approach to fuzzy quantifiers in natural languages.* Computing and mathematics with Applications, 9:149–184, 1983.

[63] Didier Dubois and Henri Prade. *Twofold fuzzy sets and rough sets - some issues in knowledge representation.* Fuzzy Sets and Systems, 23:3–18, 1987.

[64] Paul Jaccard. *Nouvelles recherches sur la distribution florale.* Bulletin de la Société de Vaud des Sciences Naturelles, pages 44–223, 1908.

[65] Amos Tversky. *Features of similarities.* Psychological Review, 84:327–352, 1977.

[66] Valerie Cross and Ted Sudkamp. *Similarity and Compatibility in Fuzzy Set Theory: assessment and applications.* Physica-Verlag, 2002.

[67] Didier Dubois and Henri Prade. *Recent Developments in Fuzzy Set and Possibility Theory*, chapter A unifying view of comparison indices in a fuzzy set-theoretic framework, pages 3–13. Pergamon Press, 1982.

[68] Bernard De Baets, Hans De Meyer, and Helga Naessens. *A class of rational cardinality-based similarity measures.* Journal of Computers and Applied Mathematics, 132(1):51–69, 2001.

[69] Bernard De Baets and Hans De Meyer. *Transitivity-preserving fuzzification schemes for cardinality-based similarity measures.* European Journal of Operational Research, 160:726–740, 2005.

[70] Didier Dubois, Philippe Fortemps, Marc Pirlot, and Henri Prade. *Leximin optimality and fuzzy set-theoretic operations.* Fuzzy Sets and Systems, 130(1):20–28, 2001.

[71] Nicolás Marín, Juan Medina, Olga Pons, D Sanchez, and Maria Villa. *Complex object comparison in a fuzzy context.* Information and Software Technology, 45:431–444, 2003.

[72] Axel Hallez, Antoon Bronselaer, and Guy De Tré. *Comparison of sets and multisets.* International Journal Of Uncertainty, Fuzziness and Knowledge-BasedSystems, 17:153 – 172, 2009.

[73] Roland Silver. *An algorithm for the assignment problem.* Communications of the ACM, 3(11):605–606, 1960.

[74] Horacio Camacho and Abdellah Salhi. *A String Metric Based on a One-To-One Greedy Matching Algorithm.* In Research in Computer Science number, pages 171–182, 2006.

[75] Tom Matthé, Rita De Caluwe, Guy De Tré, Axel Hallez, Jörg Verstraete, Marc Leman, Olmo Cornelisand Dirk Moelants, and Jos Gansemans. *Similarity Between Multi-valued Thesaurus Attributes: Theory and Application in Multimedia Systems.* In Lecture Notes in Artificial Intelligence, pages 331–342, 2006.

[76] Antoon Bronselaer and Guy De Tré. *A possibilistic approach on string comparison.* IEEE Transactions on Fuzzy Systems, 17(1):208–223, 2009.

[77] Bernard Roy. *Transitivité et connexité.* Comptes rendus de l'Académie des Sciences, 249:216–218, 1959.

[78] S Warshall. *A Theorem on Boolean matrices.* Journal of the ACM, 9(1):11–12, 1962.

[79] Esko Nuutila. *Efficient Transitive Closure Computation in Large Digraphs.* PhD thesis, Laboratory of Information Processing Science, Helsinki University of Technology, 1995.

[80] E Balas and S Yu. *Finding a maximum clique in an arbitrary graph.* SIAM Journal on Computing, 15(4):1054–1068, 1986.

[81] Vladimir Levenstein. *Binary Codes Capable of Correcting Deletions, Insertions and Reversals.* Physics Doklady, 10(8):707–710, 1966.

[82] Frederik Damerau. *A technique for computer detection and correction of spelling errors.* Communications of the ACM, 7(3):171–176, 1964.

[83] Saul Needleman and Christian Wunsch. *A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins.* Journal of Molecular Biology, 48(3):443–453, 1970.

[84] Michael Waterman, Temple Smith, and William Beyer. *Some Biological Sequence Metrics.* Advances in Mathematics, 20(4):367–387, 1976.

[85] Alvaro Monge and Charles Elkan. *The Field Matching Problem: Algorithms and Applications.* In Proceedings of the International Conference on Knowledge Discovery and Data Mining, pages 267–270, 1996.

[86] Matthew Jaro. *Unimatch: A Record Linkage System: Users Manual.* Technical report, US Bureau of the Census, 1976.

[87] William Winkler and Yves Thibaudeau. *An Application of the Fellegi-Sunter Model of Record Linkage to the 1990 US Decennial Census.* Technical Report RR91/09, Statistical Research Report Series, 1991.

[88] William Cohen, Pradeep Ravikumar, and Stephen Fienberg. *A comparison of String Distance Metrics for Name-Matching tasks.* In Proceedings of the IJCAI Workshop on Information Integration on the Web, pages 73–78, 2003.

[89] William Cohen. *Integration of Heterogeneous Databases without Common Domains Using Queries Based on Textual Similarity.* In Proceedings of the ACM Sigmod International Conference of Management of Data, pages 201–212, 1998.

[90] Mikhail Bilenko, Ray Mooney, William Cohen, Pradeep Ravikumar, and Stephen Fienberg. *Adaptive Name Matching in Information Integration.* IEEE Transactions on Intelligent Systems, 18(5):16–23, 2003.

[91] Luis Gravano, Panagiotis Ipeirotis, Nick Koudas, and Divesh Srivastava. *Text Joins in an RDBMS for Web Data Integration.* In Proceedings of the International World Wide Web Conference, pages 90–101, 2003.

[92] Robert Russell. *Soundex index*, April 1918.

[93] Howard Newcombe, James Kennedy, S Axford, and A James. *Automatic Linkage of Vital Records.* Science, 130(3381):954–959, 1959.

[94] R Taft. *Name Search Techniques.* Technical report, New York State Identification and Intelligence System, 1970.

[95] L Gill. *OX-LINK: The Oxford Medical Record Linkage System.* In Proceedings of the International Record Linkage Workshop and Exposition, pages 15–33, 1997.

[96] Lawrence Philips. *Hanging on the metaphone.* Computer Language Magazine, 7(12):39–44, 1990.

[97] Lawrence Philips. *The Double Metaphone Search Algorithm.* C/C++ Users Journal, 18(5), 2000.

[98] Gerald Salton. *Automatic text processing.* Addison-Wesley Longman Publishing Co., Inc, 1988.

[99] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern information retrieval.* ACM Press, 1999.

[100] William Cohen. *Data integration using similarity joins and a word-based information representation language.* ACM Transactions on Information Systems, 18(3):288321, 2000.

[101] Mikhail Bilenko and Raymond Mooney. *Adaptive duplicate detection using learnable string similarity measures.* In Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 39–48, 2003.

[102] Sheila Tejada, Craig Knoblock, and Steven Minton. *Learning object identification rules for information integration.* Information Systems, 26(8):607–633, 2001.

[103] Halbert Dunn. *Record Linkage.* American Journal of Public Health, 36(12):14121416, 1946.

[104] Howard Newcombe and James Kennedy. *Record Linkage: Making Maximum Use of the Discriminating Power of Identifying Information.* Communications of the ACM, 5(11):563–566, 1962.

[105] Matthew Jaro. *Advances in record linking methodology as applied to the 1985 census of Tampa Florida.* Journal of the American Statistical Society, 84(406):414–420, 1989.

[106] Herman Hartley. *Maximum likelihood estimation from incomplete data.* Biometrics, 14:174194, 1958.

[107] William Winkler. *Improved Decision Rules in the Fellegi-Sunter Model of Record Linkage.* Technical Report RR93/12, Statistical Research Report Series, 1993.

[108] William Winkler. *Methods for Record Linkage and Bayesian Networks.* Technical Report RRS2002/05, Statistical Research Report Series, 2002.

[109] N Du Bois. *A solution to the problem of Linking Multivariate Documents.* American Statistical Association Journal, 64(325):163–174, 1969.

[110] Y Wang and S Madnick. *The inter-database instance identification problem in integrating autonomous systems.* In Proceedings of the Fifth IEEE International Conference on Data Engineering, pages 46–55, 1989.

[111] Mauricio Hernandez and Salvatore Stolfo. *Real-world Data is Dirty: Data cleansing and the merge/purge problem.* Data mining and Knowledge Discovery, 2(1):9–37, 1998.

[112] M Koyuncu and A Yazici. *A fuzzy database and knowledge base environment for intelligent retrieval.* In Proceedings of the IFSA/NAFIPS World Congress, pages 2311–2316, Vancouver, Canada, 2001.

[113] Bernadette Bouchon-Meunier, Maria Rifqi, and Sylvie Bothorel. *Towards general measures of comparison of objects.* Fuzzy Sets and Systems, 84(2):143–153, 1996.

[114] Melanie Weis and Felix Naumann. *Detecting Duplicate Objects in XML Documents.* In Proceedings of the 2004 international workshop on Information quality in information systems, pages 10–19, Paris, France, 2004.

[115] Vinayak Borkar, Kaustubh Deshmukh, and Sunita Sarawagi. *Automatic segmentation of text into structured records.* ACM SIGMOD Record, 30(2):175 – 186, 2001.

[116] Antoon Bronselaer, Axel Hallez, and Guy De Tré. *Extensions of fuzzy measures and the Sugeno integral for possibilistic truth values.* International Journal of Intelligent Systems, 24(2):97–117, 2009.

[117] Andrew McCallum, Kamal Nigam, and Lyle Ungar. *Efficient clustering of high-dimensional data sets with application to reference matching.* Knowledge Discovery and Data Mining, pages 169–178, 2000.

[118] Peter Christen. *Febrl - A Freely Available Record Linkage System with a Graphical User Interface.* In Proceedings of the Second Australasian Workshop on Health Data and Knowledge Management (HDKM 2008), pages 17–25, Australia, 2008.

[119] Kathleen McKeown, Rebecca Passonneau, David Elson, Ani Nenkova, and Julia Hirschberg. *Do Summaries Help? A Task Based Evaluation of Multi Document Summarization.* In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, pages 210–217, 2005.

[120] Gerald Salton, A Wong, and C Yang. *A vector space model for automatic indexing.* Communications of the ACM, 18(11):613–620, 1975.

[121] Kanti Mardia, J Kent, and J Bibby. *Multivariate Analysis.* Academic Press, 1979.

[122] F Marriot. *Practical problems in a method of cluster analysis.* Biometrics, 27:501–514, 1971.

[123] G Milligan and M Cooper. *An examination of procedures for determining the number of clusters in a data set.* Psychometrika, 50:159–179, 1985.

[124] W Krzanowski and Y Lai. *A criterion for determining the number of groups in a data set using sum of squares clustering.* Biometrika, 44:23–34, 1985.

[125] L Kaufman and Rosseeuw. *Finding groups in data: an introduction to cluster analysis.* Wiley, 1990.

[126] Trevor Hastie, Robert Tibshirani, and Guenther Walther. *Estimating the number of clusters in a dataset via the Gap statistic.* Journal of the Royal Statistical Society, 63:411–423, 2000.

[127] Nicholas Andrews and Edward Fox. *Recent developments in document clustering.* Technical Report TR-07-35, Computer Science, Virginia Tech, 2007.

[128] George Zipf. *Selected Studies of the Principle of Relative Frequency in Language.* Harvard university press, 1932.

[129] John Hopcroft and Robert Tarjan. *Efficient algorithms for graph manipulation.* Communications of the ACM, 16:372–378, 1973.

[130] M Porter. *An algorithm for suffix stripping.* Program, 14(3):130–137, 1980.

[131] David Blei, Andrew Ng, and Michael Jordan. *Latent Dirichlet Allocation.* Journal of Machine Learning Research, 3:993–1022, 2003.

[132] Jozo Dujmovic. *A Generalization of Some Function in Continuous Mathematical Logic - Evaluation Function and its Applications.* In Proceedings of the Informatica Conference, Yugoslavia, 1973.

[133] Ronald Yager. *On ordered weighted averaging aggregation operators in multicriteria decision making.* IEEE Transactions on Systems, Man and Cybernetics, 18(1):183–190, 1988.

[134] Ronald Yager and Alexander Rybalov. *Uninorm aggregation operators.* Fuzzy Sets and Systems, 80(1):111–120, 1996.

[135] A Borgida and T Imielinski. *Decision making in committees: a framework for dealing with inconsistency and non-monotonicity.* In Proceedings Workshop of Nonmonotonic reasoning, pages 21–32, 1984.

[136] Chitta Baral, Sarit Kraus, and Jack Minker. *Combining multiple knowledge bases.* IEEE Transactions on Knowledge and Data Engineering, 3(2):208–220, 1991.

[137] Chitta Baral, Sarit Kraus, Jack Minker, and V Subrahmanian. *Combining knowledge bases consisting of first-order theories.* Computational Intelligence, 8(1):45–71, 1992.

[138] Jinxin Lin and Alberto Mendelzon. *Knowledge Base Merging by Majority.* In In Dynamic Worlds: From the Frame Problem to Knowledge Management. Kluwer, 1994.

[139] Jinxin Lin and Alberto Mendelzon. *Merging databases under constraints.* International Journal of Cooperative Information Systems, 7(1):55–76, 1998.

[140] S Konieczny and R Pérez. *Merging information under constraints: a logical framework.* Journal of Logic and Computation, 12(1):111–120, 2002.

[141] M Bright, A Hurson, and S Pakzad. *A taxonomy and current issues in multidatabase systems.* Computer, 25(3):50–59, 1992.

[142] Sandra Sandri, Didier Dubois, and Henk Kalfsbeek. *Elicitation, assessment and pooling of expert judgements using possibility theory.* IEEE Transactions on Fuzzy Systems, 3(3):313–335, 1995.

[143] Sebastien Destercke, Didier Dubois, and Eric Chojnacki. *Possibilistic Information Fusion using Maximal Coherent Subsets.* IEEE Transactions on Fuzzy Systems, 17(1):79–92, 2009.

[144] Axel Hallez, Guy De Tré, Jörg Verstraete, and Tom Matthé. *Application of fuzzy quantifiers on possibilistic truth values.* In Proceedings of the Eurofuse Workshop on Data and Knowledge Engineering, pages 252–254, Warshau, Polen, 2004.