godk.datum vitsord

bedömare

# Perspektiv på språkprocessering: Disambiguering av ord och automatisk sammanfattning

Fabian Fagerholm

Helsingfors den 9 maj 2005

HELSINGFORS UNIVERSITET
Institutionen för datavetenskap

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

| Tiedekunta/Osasto — Fakultet/Sektion — Faculty | Laitos — Institution — Department |
|---|---|
| Matematisk-naturvetenskapliga | Datavetenskapliga institutionen |

Tekijä — Författare — Author
Fabian Fagerholm

Työn nimi — Arbetets titel — Title

Perspektiv på språkprocessering: Disambiguering av ord och automatisk sammanfattning

Oppiaine — Läroämne — Subject
Datavetenskap

| Työn laji — Arbetets art — Level | Aika — Datum — Month and year | Sivumäärä — Sidoantal — Number of pages |
|---|---|---|
| Avhandling | den 9 maj 2005 | 24 sidor + 3 bilagesidor |

Tiivistelmä — Referat — Abstract

Automatisk språkprocessering har efter mer än ett halvt sekel av forskning blivit ett mycket viktigt område inom datavetenskapen. Flera vetenskapligt viktiga problem har lösts och praktiska applikationer har nått programvarumarknaden.

Disambiguering av ord innebär att hitta rätt betydelse för ett mångtydigt ord. Sammanhanget, de omkringliggande orden och kunskap om ämnesområdet är faktorer som kan användas för att disambiguera ett ord.

Automatisk sammanfattning innebär att förkorta en text utan att den relevanta informationen går förlorad. Relevanta meningar kan plockas ur texten, eller så kan en ny, kortare text genereras på basen av fakta i den ursprungliga texten.

Avhandlingen ger en allmän översikt och kort historik av språkprocesseringen och jämför några metoder för disambiguering av ord och automatisk sammanfattning. Problemområdenas likheter och skillnader lyfts fram och metodernas ställning inom datavetenskapen belyses.

ACM Computing Classification System (CCS):
I.2.7 [Natural Language Processing],
H.3.1 [Content Analysis and Indexing],
H.3.3 [Information Search and Retrieval]: Information filtering

Avainsanat — Nyckelord — Keywords
språkprocessering, disambiguering av ord, automatisk sammanfattning

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — övriga uppgifter — Additional information

# Innehåll

# 1 Inledning

Datavetenskapliga applikationer har möjliggjort ett enormt flöde av språklig information i samhället. Mängden överstiger i många fall den mänskliga kapaciteten, och ett behov av automatisk språkprocessering har uppkommit. Området har varit föremål för forskning i över 50 år, och kan sägas vara ett av de ämnen som gått hand i hand med datavetenskapen sedan dess början. Men faktum är att språkprocesseringen genom sin historia inte bara har gett spännande och användbara resultat, utan den har också stött på stora besvikelser och motgångar.

Språkprocesseringen har gått från betraktelse av enskilda ord till betraktelse av hela texter, och de tidiga resultaten har främst handlat om att betrakta vad texten har sagt ordagrant. Texten har undersökts på ytan och av praktiska skäl har den yttre kunskap om ämnet som behövs för en avancerad tolkning lämnats bort. Småningom börjar dock forskningen nå en punkt där man kan betrakta också de implicita betydelserna. Perspektivet på språk växlar mellan en ordagrann tolkning och en tolkning av språket som ett symbolsystem där sammanhanget och yttre kunskap avgör vilka värderingar som gäller [HC81, Faw92].

I denna avhandling behandlar jag två perspektiv på språkprocessering, disambiguering av ord och automatisk sammanfattning. Jag presenterar några metoder som används för dessa, och visar på vilket sätt metoderna liknar och skiljer sig från varandra. Jag visar också hur disambiguering av ord och automatisk sammanfattning relaterar till varandra. Samtidigt anknyter jag till deras ställning inom språkprocesseringen och datavetenskapen.

# 2 En översikt av språkprocessering

Språkprocessering ingår som ett element inom många områden av datavetenskapen. Samtidigt ingår många områden av datavetenskapen i språkprocesseringen. Beroende på infallsvinkel kan språkprocesseringen alltså sägas vara ett delproblem av andra datavetenskapliga problem, eller så kan de andra problemen vara byggstenar i språkprocesseringen. Till exempel artificiell intelligens, maskininlärning, informationssökning och användargränssnitt är områden som innehåller språkprocessering som ett element, men som samtidigt har producerat resultat som språkprocesseringen använder sig av.

En infallsvinkel är att förankra språkprocesseringen i dess applikationer. I tabell 1

| | **Färdiga resultat** | **Utvecklingsstadium 1995** | **Framtidsvision** |
|---|---|---|---|
| **Ordbehandling** | Enkel sökning av textsträngar | Stavningskontroll | Grammatikkontroll |
| **Maskinöversättning** | Översättning med glosor | Översättningsminnen, direkt mapping | Flerspråkig översättning |
| **Processering av naturligt språk** | Språkgränssnitt | Generering ur databas, extraktion av entiteter | Extraktion av händelser |
| **Informationssökning** | Sökning med nyckelord | Sökning med naturligt språk | Sökning på basen av koncept |

Tabell 1: Utveckling av problem inom språkprocesseringen; adapterad från "Commercial Applications of Natural Language Processing" [CR95]. Tabellen visar utvecklingen av fyra olika problemområden (en rad per område). I tabellen löper tiden från vänster till höger, och situationen 1995 är i mitten. På vissa områden har utvecklingen under de senaste tio åren varit snabb, t.ex. grammatikkontroll och sökning på basen av koncept är redan status quo.

visas utvecklingen hos fyra av språkprocesseringens problemområden: ordbehandling, maskinöversättning, processering av naturligt språk och informationssökning. Tabellen är från 1995 och en del av framtidsvisionerna är redan en realitet efter tio år av forskning. I resten av detta kapitel kommer vi att gå igenom de olika områdena som visas i tabellen och se hur utvecklingen har gått framåt.

## 2.1 Språkprocesseringens historiska utveckling

Språkprocessering var av politiska skäl ett viktigt område under 1950- och 1960-talen. I USA fanns ett intresse för dokument skrivna på ryska. Den amerikanska försvarsmakten och underrättelsetjänsten reserverade medel för forskning i hopp om att kunna automatisera språkliga uppgifter som dittills hade utförts för hand. Under mitten av 1960-talet började man dock ifrågasätta behovet av att bedriva dyr forskning som inte såg ut att leda till konkreta resultat. Finansiärerna tillsatte en kommitté för att utreda saken.

År 1966 redovisade ALPAC-kommittén (Automatic Language Processing Advisory Committee) sin rapport. Rapporten behandlade många aspekter av den dåvarande forskningen, och gav rekommendationer för fortsatta forskningsprojekt. Rapporten tog också ställning till ekonomiska frågor. Den förutspådde att ordbehandling skulle bli ett viktigt kommersiellt område [CR95]. Idag är det en självklarhet att ordbehandlingsprogram har åsidosatt den traditionella skrivmaskinen. Ordbehandlingen

har fortsatt att utvecklas (se tabell 1, rad 1), och gått från en avancerad skrivmaskin till att innehålla funktioner som har med själva språket att göra. Idag är ordbehandlaren också en språkbehandlare – här syns tydligt steget från ord till språk.

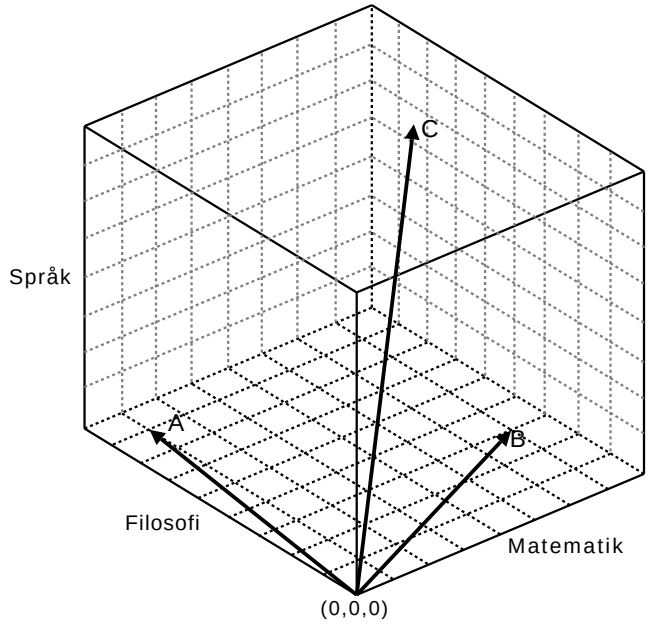
Maskinöversättningen hör till språkprocesseringens tidiga misslyckanden (tabell 1, rad 2). ALPAC-rapporten visade att förväntningarna på maskinöversättning under mitten av 1900-talet hade varit för höga [CR95]. De enkla metoder där man översätter ord för ord med hjälp av glosor ledde inte till tillfredsställande resultat, även om metoden i sig var enkel. ALPAC-rapporten konstaterade att traditionell översättning var ett förhållandevis förmånligt alternativ. Maskinöversättningen fick ett dåligt rykte och många av forskningsprojekten avslutades.

Småningom har dock olika tekniker tagits fram som producerar långt bättre resultat än de tidiga försök som ALPAC-rapporten kunde evaluera. Översättningsminnen är en teknik där man sparar de översatta meningarna för återanvändning i senare översättningar. I direkt mapping gör man ett semantiskt träd av satsstrukturen i den ursprungliga texten, och omvandlar trädet så att det bättre motsvarar målspråket. Först därefter gör man den egentliga översättningen.

Processering av naturligt språk är det största av de fyra problemområdena i tabell 1, och innehåller i sig ett stort antal mindre problem. Man kan säga att området har förgrenats från informationssökningen – naturliga sökfraser måste brytas ner till en semantisk representation för att de skall gå att använda som sökkriterier. Språkgränssnittet handlar om hur man kan bryta ut och identifiera ord och meningar i text eller ljud, och göra grundläggande grammatikalisk analys av de individuella fragmenten. Följande steg är att identifiera entiteter, till exempel namn, telefonnummer eller mer allmänt *vem* gjorde *vad* åt *vem/vilka*. Denna information kan sedan användas för att till exempel ge kommandon åt datamaskinen eller konstruera en logisk modell av det som texten beskriver. Det omvända kallas textgenerering; då går man från strukturerad data i en databas till en beskrivning i naturligt språk. Följande utvecklingsnivå betraktar kausalförhållanden i texten och handlar alltså till exempel om att kunna ordna händelser i kronologisk ordning trots att de i texten kan förekomma i en annan ordning [CR95].

## 2.2 Vektorrymden – en grundläggande modell

Inom informationssökningen har utvecklingen börjat vid enkel matchning av nyckelord, och kombinationer av nyckelord genom boolsk algebra (tabell 1, rad 4). Små-




Figur 1: I vektorrymdsmodellen representeras attributen i ett dokument som element av en vektor. Att jämföra två dokument är liktydigt med att jämföra vektorerna som representerar dem. Här visas en fiktiv tredimensionell rymd. Varje dimension beskriver ett attribut, och i rymden visas tre fiktiva dokument. Vektorerna visar i hurdan grad dokumenten hör ihop med de ämnesord som attributen representerar.

ningom har man utvecklat system där användaren kan mata in sökkriteriet i naturligt språk, och systemet konverterar kriteriet till en mera formell form. Följande steg har igen gått från ord till en högre nivå, och de mest sofistikerade sökteknikerna arbetar på konceptnivå. Ett exempel är Latent Semantic Analysis (LSA) [DDL$^+$90].

LSA bygger på *vektorrymdsmodellen* som är populär inom informationssökningen. I modellen representeras varje dokument som en vektor, och vektorns element beskriver valda attribut för dokumentet. Attributen kan vara binära, och representerar då existens eller avsaknad av ifrågavarande attribut, eller vikter, och representerar då i vilken grad varje attribut förekommer i dokumentet. Alternativt kan vikten representera hur viktigt attributet är i just det dokumentet.

I figur 1 visas en fiktiv vektorrymd. Den är tredimensionell, och varje dimension representerar ett ämnesord eller -område. I rymden finns tre vektorer som var och en representerar ett dokument. Vektorn $A$ kunde representera Bertrand Russels grundläggande filosofiska verk "Västerlandets filosofi", vektorn $B$ en lärobok i statistik, och vektorn $C$ denna avhandling. Vektorerna visar i hurdan grad dokumentet hör

ihop med vart och ett av ämnesorden. Om man vill jämföra två dokument, kan man alltså jämföra vektorerna som representerar dokumenten. Ju mer vektorernas längd och riktning i alla dimensioner liknar varandra, desto mer liknar dokumenten varandra. Man kan också omvandla ett sökkriterium till en vektor och hitta relevanta dokument med samma sorts jämförelse [JM00].

För att praktiskt kunna använda vektorrymdsmodellen måste man utföra någon förhandsprocessering av vektorerna, annars kan irrelevanta faktorer som dokumentets längd störa resultatet. LSA definierar en metod för att använda vektorrymdsmodellen till att hitta en latent betydelse i dokumentet, men där inverkar ändå dokumentets längd på resultatet. Vidare utveckling av LSA har åtgärdat det problemet [JM00, DDL+90].

Inom automatisk sammanfattning och disambiguering av ord är vektorrymdsmodellen i en central ställning, både på grund av sin solida matematiska definition och det faktum att dagens datorer är väl anpassade för matrisräkning.

# 3 Disambiguering av ord

Disambiguering av ord innebär att hitta rätt betydelse av ett mångtydigt ord. Som ett enkelt exempel ser vi att ordet “rätt” i den förra meningen kan betyda både “korrekt” eller “mat” beroende på sammanhanget. Lösryckt är det svårt att veta vad som avses med ordet.

Något förenklat kan man tänka sig att problemet består av två delproblem: att fastställa alla möjliga betydelser hos ett ord, och att välja vilken betydelse ett visst ord har i ett specifikt fall. Uppgiften är ofta enkel för en människa, men det visar sig att den är mycket svår att lösa algoritmiskt. För att fullständigt lösa problemet med disambiguering av ord, krävs en lösning på alla svåra problem inom artificiell intelligens, till exempel representation av sunt förnuft och encyclopedisk kunskap. Språkets mångtydiga natur är fundamentalt annorlunda än den matematiska exakthet som kännetecknar datavetenskapen.

Disambiguering av ord är ett mellanliggande mål som behövs för att lösa flera andra problem inom språkprocesseringen [IV98]. Inom maskinöversättning är det till exempel ofta nödvändigt att välja rätt betydelse för ett ord innan man kan hitta en motsvarighet i ett annat språk. Informationssökning kräver också ofta att man kan skilja på mångtydiga söktermer och välja rätt betydelse bland de olika alternativen.

Det finns olika lösningsstrategier för de två delproblemen ovan. Det första delproblemet, att fastställa alla betydelser hos ett ord, har ofta lösts genom att för hand göra upp en lista på alla betydelser för alla ord. Det andra delproblemet, att välja rätt betydelse, har lösts antingen genom att betrakta enbart de omkringliggande orden, eller genom att dessutom utnyttja kunskapskällor utanför texten.

Alla system för disambiguering av ord kan jämföras med en enkel metod: att välja den mest frekventa betydelsen. Den mest frekventa betydelsen fås genom att hämta alla förekomster av det mångtydiga ordet ur ett stort textmaterial, disambiguera varje fall för hand, och sedan räkna ut vilken betydelse som är den vanligaste. En metod som inte ger bättre resultat än denna har liten praktisk betydelse. Man får inte glömma att textmaterialet kan påverka betydelsen, och då man evaluerar olika metoder för disambiguering är det vanligt att man använder något standardmaterial för att få jämförbara resultat.

Bristen på maskinläsbara ordlistor var länge en av de största begränsningarna för disambiguering. Idag har situationen avhjälpts för det engelska språket med bl.a. WordNet [Fel98], som på grund av sin tillgänglighet för allmänheten har haft stor inverkan på området [IV98]. WordNet är både en ordbok med förklaringar av ordens betydelser, och en nätverksstruktur som beskriver relationerna mellan ord, till exempel synonymer och begreppshierarkier.

För att illustrera betydelsen av WordNet skall vi betrakta ett exempel på ett mångtydigt ord:

> His honour had further observed, that a female Yahoo would often stand behind a **bank** or a bush, to gaze on the young males passing by, and then appear, and hide, using many antic gestures and grimaces, at which time it was observed that she had a most offensive smell; and when any of the males advanced, would slowly retire, looking often back, and with a counterfeit show of fear, run off into some convenient place, where she knew the male would follow her [Swi97].

I detta stycke ur Jonathan Swifts "Gullivers resor" förekommer det mångtydiga engelska ordet *bank*, som jag har markerat. I WordNet hittar vi följande förklaringar på ordet:

> The noun bank has 10 senses (first 9 from tagged texts)
>
> 1. (883) depository financial institution, bank, banking concern, banking company – (a financial institution that accepts deposits and channels

the money into lending activities; “he cashed a check at the bank”; “that bank holds the mortgage on my home”)

2. (99) bank – (sloping land (especially the slope beside a body of water); “they pulled the canoe up on the bank”; “he sat on the bank of the river and watched the currents”)
3. (76) bank – (a supply or stock held in reserve for future use (especially in emergencies))
4. (54) bank, bank building – (a building in which commercial banking is transacted; “the bank is on the corner of Nassau and Witherspoon”)
5. (7) bank – (an arrangement of similar objects in a row or in tiers; “he operated a bank of switches”)
6. (6) savings bank, coin bank, money box, bank – (a container (usually with a slot in the top) for keeping money at home; “the coin bank was empty”)
7. (3) bank – (a long ridge or pile; “a huge bank of earth”)
8. (1) bank – (the funds held by a gambling house or the dealer in some gambling games; “he tried to break the bank at Monte Carlo”)
9. (1) bank, cant, camber – (a slope in the turn of a road or track; the outside is higher than the inside in order to reduce the effects of centrifugal force)
10. bank – (a flight maneuver; aircraft tips laterally about its longitudinal axis (especially in turning); “the plane went into a steep bank”)
[Fel98]

Ordet har dessutom åtta betydelser som verb, men vi skall här betrakta bara dessa substantivbetydelser. Siffrorna inom parentes berättar hur många gånger ordet har förekommit i källtexterna till WordNet – texter som har använts för att manuellt disambiguera ord och lägga in dem i WordNet-databasen.

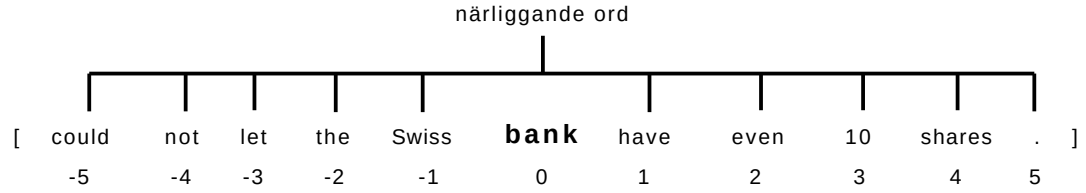
För en människa med kunskap om det engelska språket är det inte svårt att se att ordet *bank* i Swifts text står i den sjunde betydelsen: “a long ridge or pile”. Vi skall nu betrakta några algoritmiska sätt att disambiguera ordet.

## 3.1 Disambiguering med hjälp av kontext

Genom att betrakta de omkringliggande orden kan man nå förvånansvärt goda resultat. En central fråga är hur många omkringliggande ord som i regel behövs för att disambiguera ett ord. Man kan välja hela meningen, eller så kan man ta ett förutbestämt antal ord. En annan central fråga i hela språkprocesseringen är vad som definieras som ett ord: är till exempel kommatecken och punkt ord i sig? Svaret beror på sammanhanget och användningsområdet.

| | | |
|---|---|---|
| the state to take over | **bank** | accounts, stocks and other |
| and undermine the confidence of | **bank** | customers. “If you |
| occupied all of the left | **bank** | of the Rhine. The |
| . The wife of convicted | **bank** | robber Lawrence G. Huntley was |
| could not let the Swiss | **bank** | have even 10 shares. |
| City, N.J. That | **bank** | handles most of the paper |
| sharp bows into the soft | **bank**. | The flat-bottomed boat swung |
| shareholders. I visited the | **bank** | in March and wrote a |
| , the people at the | **bank** | said they felt that they |
| get in touch with the | **bank**. | Doyle cannot undertake to |
| boat financed through a local | **bank** | is done much the same |
| , credit rating and local | **bank** | policy. Outboard motors , |
| shop, or make a | **bank** | deposit, the ever-increasing number |
| Newbury shore and the south | **bank** | of Deer Island there was |
| Waco down through the cloud | **bank** | and hope to break through |
| of a representative of the | **bank** | that held the papers against |
| end, and that the | **bank** | would accept a mortgage on |
| fight from behind the arroyo | **bank**. | Bullets were so thick |
| a plant on the west | **bank** | of the Battenkill south of |
| did seem impossible. The | **bank** | which held the mortgage on |
| to crawl back up the | **bank** | toward the rampart. Watson |
| . Watson stumbled down the | **bank**. | The man leaned his |
| off, up the river | **bank**. | He wanted a careful |

Figur 2: Ett fragment ur en konkordans. Källmaterialet är Brown-korpusen och här betraktas tio omkringliggande ord, fem till vänster och fem till höger. Meningarna är medvetet valda för att illustrera olika betydelser hos ordet *bank*. I vissa fall är det svårt till och med för en människa att med säkerhet avgöra vilken betydelse som är den rätta med så här lite kontext. Notera att skiljetecken har räknats som skilda ord.




Figur 3: En kollokation eller ordförbindelse beaktar vilka ord som finns i specifika positioner kring ett mångtydigt ord.

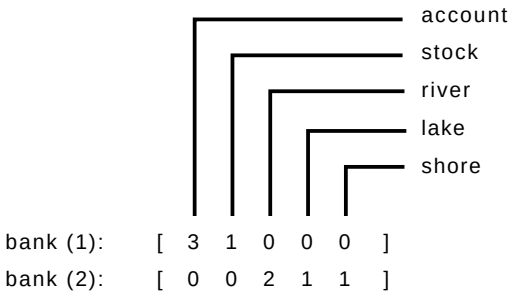
Disambigueringsprocessen börjar med att skapa en så kallad *konkordans* av ett källmaterial. Konkordansen för ett visst ord innehåller textfragment där ordet omges av något antal omkringliggande ord. I figur 2 visas ett litet stycke av en konkordans vars källmaterial är den så kallade Brown-korpusen [FK64], en samling texter ur amerikansk litteratur. Konkordansen betraktar de närmaste tio orden kring ordet *bank*, fem ord före och fem ord efter. I detta fall har skiljetecken betraktats som egna ord.

Ur konkordansen kan man hämta olika slags information om det mångtydiga ordet. Denna information fungerar som indata för olika disambigueringsmetoder. I huvudsak kan man leta efter två aspekter:

- Kollokationer eller ordförbindelser, och
- samförekomst av ord.

Då man letar efter ordförbindelser beaktar man vilka ord som finns i de olika positionerna kring det mångtydiga ordet. Ofta tar man också med information om de olika ordens grundform eller ordstam och grammatikaliska ställning i satsen, till exempel subjekt, predikat och andra satsdelar. Observationerna kodas i vektorform, och dessa vektorer kan ofta användas för att effektivt välja rätt betydelse för det mångtydiga ordet. I figur 3 visas ett enkelt exempel utan grammatikalisk information. Man kan tänka sig att förekomsten av ordet *Swiss* i position -1 ofta leder till att ordet *bank* har den första betydelsen (“depository financial institution”) i WordNet.

Å andra sidan kan man analysera de närmast omkringliggande orden utan att beakta deras exakta position till vänster eller höger om det mångtydiga ordet. Då blir varje närliggande ord ett element i vektorn, och elementets värde är antalet förekomster av respektive ord i närheten av det mångtydiga ordet i hela källmaterialet. Figur 4 är ett exempel på detta. Det är klart att ett stort textmaterial ger upphov till




Figur 4: Samförekomst av ord beaktar hur ofta vissa betydelsefulla ord förekommer i samband med ett visst mångtydigt ord i hela källmaterialet. Detta exempel visar några betydelsefulla ord som kunde förekomma i samband med två olika betydelser av ordet *bank*. Betydelserna följer uppdelningen i WordNet.

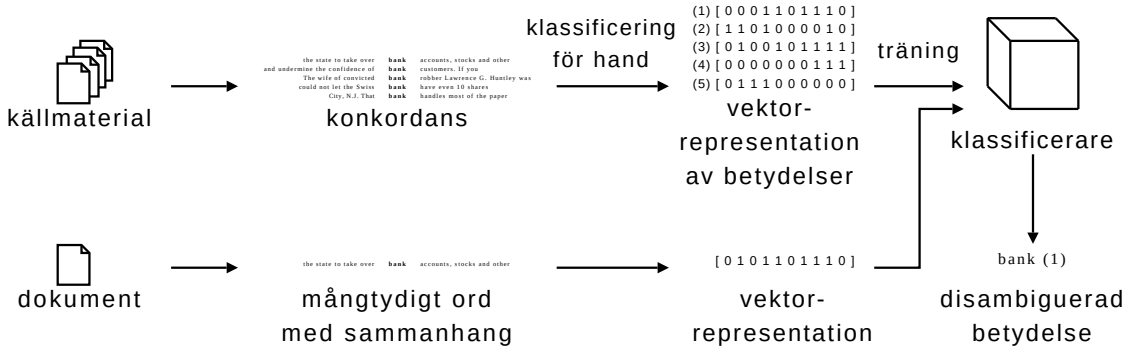
vektorer av mycket hög dimensionalitet, i vilka de flesta dimensioner inte har någon disambiguerande effekt. Därför filtrerar man ofta bort mycket frekventa ord och lämnar kvar bara sådana ord som verkligen differentierar olika betydelser [JM00]. Denna förhandsprocessering är ett problemområde i sig.

Då man har samlat information om tillräckligt många ord och lagrat informationen i vektorform, kan man använda vektorerna för att disambiguera ord i andra texter. Det finns flera olika sätt att göra detta på, men ett gemensamt drag hos de flesta metoderna är att de bygger på statistik och sannolikhetslära. Grundtanken är att man skapar en klassificerare som kan tränas med vektorinformationen som fåtts ur källmaterialet. Klassificeraren kan sedan välja rätt klass, alltså rätt betydelse, för vektorer ur nya texter. Till valet hör ofta ett sannolikhetsvärde som anger hur sannolikt det är att ordet har den valda betydelsen. I figur 5 visas hur de olika stegen i disambigueringsprocessen kopplas ihop.

## 3.2 Disambiguering med yttre kunskap

Yttre kunskap, eller kunskap om den "värld" som texten behandlar, är en annan källa som kan användas för att disambiguera ord. Dessa metoder kommer närmare artificiell intelligens än de rent statistiska metoder som används för disambiguering med hjälp av kontext. Målet är att kunna resonera om orden och efterlikna människans sätt att väga de olika alternativen.

Det finns vissa enkla metoder för att underlätta disambigueringsproblemet då man har tillgång till grundläggande yttre kunskap. De grundar sig ofta på observationer




Figur 5: Disambigueringsprocessen består av flera olika steg. I disambiguering som bygger på en korpus med förklassificerade dokument börjar processen för varje ord med att ordets alla förekomster extraheras ur källmaterialet och en konkordans med förekomsterna och något omkringliggande material skapas. Varje förekomst disambigueras för hand och den valda betydelsen lagras i vektorform tillsammans med information om det omkringliggande materialet. Med hjälp av vektormaterialet tränas en klassificerare. Senare kan klassificeraren användas för att välja en betydelse för nya ord som hämtas ur andra texter och omvandlas till vektorform.

om vissa typer av text. Om man till exempel vet att texten är en ekonomisk nyhetstext, så kan man genast dra slutsatsen att vissa betydelser för ord som *bank* är mycket mindre sannolika än andra betydelser. Ämnesområdet medför således vissa restriktioner som kan reducera antalet alternativ [IV98].

Att nå goda resultat för allmän text med disambiguering med yttre kunskap har dock visat sig vara ytterst svårt. Dels är ämnesområdet inte begränsat, och dels har man ännu kvar en stor mängd komplicerade problem. De statistiska metoderna har lett till praktiska resultat, och har därför gått före under 80- och 90-talen. Småningom har olika metoder för att lösa delproblem inom disambiguering med yttre kunskap utarbetats, och vissa forskare har återgått till tidigare försök att algoritmiskt bilda en så kallad djup förståelse av texten.

I dessa metoder närmar man sig ofta problemet genom att skapa en semantisk hierarki av ord och begrepp. Den semantiska hierarkin representerar ett sorts sunt förnuft enligt vilket man kan resonera om något ämnesområde. Målet är att kunna utesluta de betydelser som inte är förnuftiga. En datastruktur som är konstruerad med tanke på att kunna resonera om objekt och deras relationer inom ett visst ämnesområde kallas en *ontologi*. Man kan utesluta vissa betydelser genom att för närliggande ord hitta ett gemensamt begrepp på en högre nivå [IV98].

Även om WordNet inte från början konstruerades med tanke på detta, har denna datastruktur fått egenskaper som liknar en ontologi. Den semantiska hierarkin i WordNet har också använts för att disambiguera ord. WordNet lagrar så kallade hypernymrelationer, som är av formen *x är en sorts y*. I exemplet från "Gullivers resor" (sidan 6) där ordet *bank* skulle disambigueras, kunde man tänka sig att följa hierarkin i WordNet och hitta ett gemensamt, mer allmänt koncept på högre nivå med ordet *bush*. Då kan man för det första utesluta alla betydelser som inte hör till denna gemensamma övre kategori, och för det andra välja den betydelse som har det kortaste avståndet till ordet *bank*.

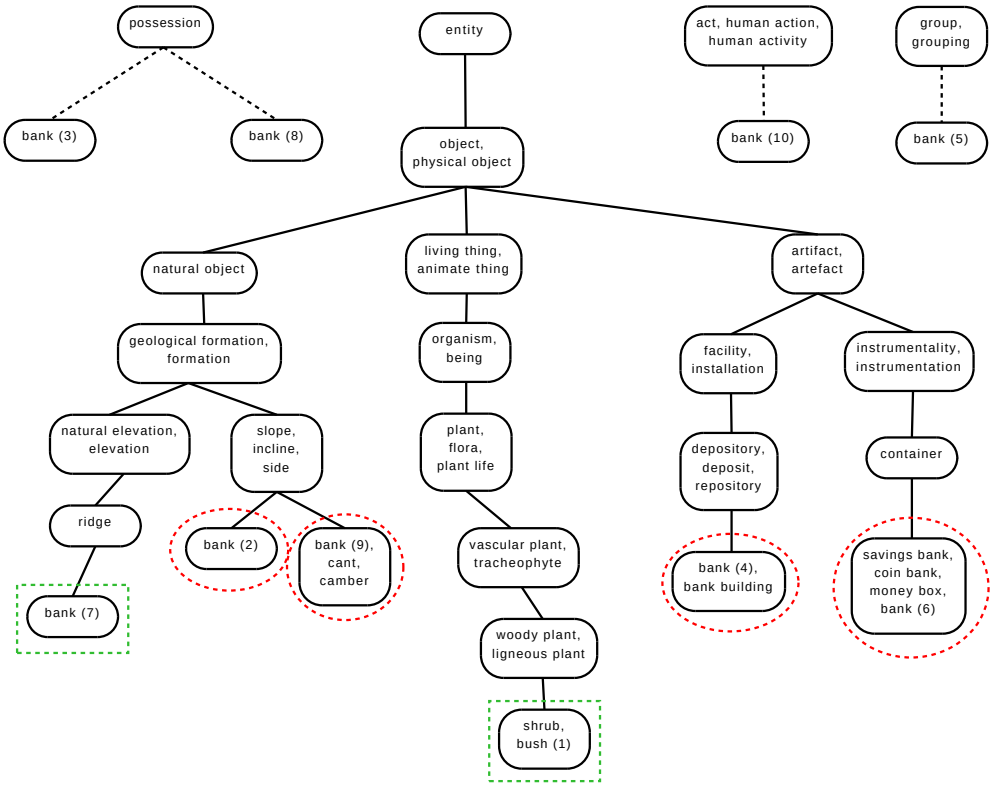
I figur 6 visas ett fragment av hypernymgrafen för några av betydelserna hos orden *bank* och *bush*. Vi antar här att det senare ordet redan har disambiguerats på annat sätt, och därför kan vi använda det som referenspunkt.

Nackdelen med WordNet syns här; ordet kan inte helt disambigueras eftersom fyra betydelser är lika nära *bush*. Dessutom är den rätta betydelsen inte den närmaste. Hypernymrelationerna i WordNet räcker inte till för att i detta fall hitta den rätta betydelsen, men här har dock fyra betydelser helt kunnat uteslutas. Genom att betrakta andra relationer i WordNet kan man möjligen utesluta ännu fler. Några exempel är meronymer (*x består av a, b, ...*), holonymer (*x är en del av a, b, ...*) och hyponymer (*a, b, ... är en sorts x*).

## 3.3 Praktiska implementeringar

I praktiken används metoder som kombinerar olika aspekter av de teoretiska modellerna. Ett praktiskt exempel på en metod som lånar idéer från bland annat disambiguering med yttre kunskap, beskrivs av Jen Nan Chen och Jason Chang [CC98]. De observerar att metoder som bygger på träning genom förklassificerade dokument sannolikt gör att disambigueringen bara fungerar bra i specifika textsorter som liknar träningsmaterialet. För praktiska ändamål är det viktigt att metoden kan klassificera ett stort antal texter, och dessutom att den är flexibel nog att klara av helt nya textsorter.

Chen och Chang föreslår en metod där man gör två steg av disambiguering och lagrar inte bara de närliggande orden utan också konceptkategorier i samförekomstvektorerna. I det första steget disambigueras alla de ord som kan disambigueras med hjälp av en maskinläsbar ordlista och en tesaurus. Systemet ger ett sannolikhetsvärde för varje disambiguering, och sådana disambigueringar vars sannolikhetsvärde




Figur 6: I WordNet [Fel98] beskrivs bl.a. hypernymrelationerna mellan ord. Hypernymer betecknar relationer av formen *x är en sorts y*. Genom att hitta ett gemensamt övre begrepp för två betydelser av två närliggande ord och välja det par vars avstånd i grafen är kortast, kan man i bästa fall helt disambiguera ett ord. Här är dock den rätta betydelsen av ordet *bank* (7) ett steg längre ifrån ordet *bush* än fyra andra betydelser (2, 9, 4 och 6), som alla är på samma avstånd, så disambigueringen blir inte korrekt eller entydig enbart genom denna metod. Några helt orelaterade betydelser har dock kunnat uteslutas (3, 8, 10 och 5). I figuren är den korrekta relationen markerad med en grön rektangel, och de inkorrekta betydelserna av *bank* som inte har kunnat uteslutas är markerade med en röd cirkel.

är för lågt, förkastas som inkorrekta. Därefter görs en ny konkordans av de korrekt disambiguerade orden, och denna matas in i systemets disambigueringsdatabas eller klassificerare. Då kommer databasen att innehålla information om vilka ord som tidigare har förekommit i närheten av andra ord, och också vilka koncept eller betydelser de då har haft. Slutligen görs en andra disambiguering och ett större antal ord kan nu disambigueras. Chen och Chang visar att en dylik metod är realistisk, att den leder till relativt goda resultat och dessutom att den är helt automatisk i alla skeden [CC98].

# 4 Automatisk sammanfattning

Ett problemområde där disambiguering av ord kan användas är automatisk textsammanfattning. Syftet med automatisk sammanfattning är att omvandla en lång text till en kortare text, utan att lämna bort det relevanta. Jämfört med disambiguering av ord kan detta vara en utmaning till och med för en människa.

De flesta metoder för automatisk sammanfattning tillhör en av två kategorier: textutdrag eller faktautvinning. Textutdrag går ut på att plocka ut fragment av den ursprungliga texten och sätta ihop dem till en ny text. Faktautvinning går ut på att leta efter en viss sorts information i den ursprungliga texten, göra en semantisk struktur av den, och slutligen generera en ny text på basen av strukturen [SJ98].

Man kan säga att textutdrag motsvarar valda citat ur en text – ett sammandrag av de viktigaste meningarna – medan faktautvinning motsvarar ett abstrakt eller en regelrätt sammanfattning där den viktigaste informationen upprepas i en ny formulering. I denna avhandling används termen sammanfattning för att täcka hela området, medan termen sammandrag syftar på enbart den första betydelsen.

## 4.1 Textutdrag

Ett sätt att producera ett sammandrag av en text är att ta bort de flesta meningarna och bara lämna kvar de meningar som är mest relevanta. Då reduceras problemet till att hitta de relevanta meningarna, och man behöver inte bry sig om att producera någon ny text. Syftet är inte nödvändigtvis att producera en högklassig återgivning av det ursprungliga materialet, utan snarare att göra det lättare att sålla fram relevanta dokument ur ett stort material [Zec95].

Ett enkelt sätt att producera ett sammandrag är att välja de $n$ första och $m$ sista

meningarna i en text, eller att välja ett antal meningar slumpmässigt. Om en algoritm för textutdrag skall ha något värde, måste den producera ett sammandrag som är bättre än resultatet av en sådan enkel metod.

Textutdrag har mycket gemensamt med metoderna för disambiguering med hjälp av kontext. Liksom att representera betydelsen hos varje ord som en relation mellan ordet och omkringliggande ord, representeras varje mening som en relation mellan meningen och dess ord. Det visar sig att en sådan metod i det allmänna fallet bättre väljer ut ett antal relevanta meningar än en där början och slutet av texten bildar ett sammandrag.

Genom att poängsätta varje mening kan man enkelt välja ut ett givet antal av de mest relevanta meningarna. Poängsättningen kan bestå av flera olika mätningar som alla bidrar till en poängsumma för meningarna. Före poängsättningen filtreras mycket högfrekventa och redundanta ord bort (till exempel “och”, “för”, “med”), och orden kan omvandlas till ordstammar genom att klippa bort ändelser. Därefter kan poängsättningen börja. I flera experiment har bland annat följande mätningar eller variationer av dem använts [Zec95]:

- Nyckelord: vissa nyckelord inverkar positivt eller negativt på poängsättningen.
- Rubrikord: ord i rubriker och underrubriker väljs som nyckelord, och meningarna får poäng på basen av hur många av dessa ord de innehåller.
- Plats: meningar i början och slutet av texten, och i början och slutet av stycken, ges högre poäng.
- Längdkriterium: korta meningar lämnas bort, eller meningarna får poäng på basen av längd – för korta eller för långa meningar får färre poäng.
- Fraser: om en mening innehåller samma fras som en rubrik, får den högre poäng.
- Temaord: de mest frekventa orden väljs som temaord, och meningarna får fler poäng ju fler temaord de innehåller.

Variationer på dessa kan hittas inte bara i metoder för automatisk sammanfattning, utan också i databrytning ur text och informationssökning.

Även om platskriteriet i allmänhet inte producerar goda sammandrag, kan det med modifikation vara lämpligt för text där informationen typiskt är upplagd på ett speciellt sätt. I nyhetstext strävar man till att placera den mest relevanta informationen

först, och sedan utvidga detaljerna allteftersom relevansen sjunker. I teorin betyder det att en nyhetstext skall gå att klippa av var som helst, och den återstående delen ovanför klippet skall ändå gå att förstå. I praktiken fungerar detta bara på mycket kort telegram- eller rapporttext – i andra slag av nyhetstext är det ändå ofta så att den relevanta informationen finns spridd över hela texten.

SweSum är ett verktyg för automatisk sammanfattning av nyhets- och rapporttext, utvecklat av Martin Hassel och Hercules Dalianis vid Kungliga Tekniska Högskolan (KTH) i Stockholm [Dal00]. SweSum beaktar bland annat platsinformation genom att ge varje mening ett poängvärde $1/n$ där $n$ är meningens ordningstal. Dessutom ges fetstil och meningar med numerisk information ett högre poängvärde.

SweSum använder förutom platsinformation också poängsättning med hjälp av nyckelord. Systemet omvandlar orden till ordstammar och poängsätter varje mening enligt nyckelordsfrekvens. Nyckelorden tas från en databas med ca 700 000 ord, och dessutom kan användaren mata in egna nyckelord för att styra sammanfattningen.

I bilaga A visas ett exempel på ett sammandrag producerat av SweSum. Ett problem med sammandraget är att koherensen har blivit lidande; en mening vars betydelse är beroende av en annan mening har ensam tagits med i sammandraget, och blir därför missvisande. Dalianis har i samarbete med studenter vid KTH undersökt just denna aspekt av SweSum och i specifika exempelfall kommit till vissa gränsvärden på sammandragets längd vid vilka koherensen blir lidande [Dal00].

## 4.2 Faktautvinning

Textutdrag rör sig på ytnivån av texten. Faktautvinning rör sig på en semantisk nivå och försöker skapa en struktur av den information som finns i texten. Genom att välja lämpliga bitar av denna struktur kan man utvinna fakta för en ny text [SJ98, Zec95].

Strukturen kan vara av olika slag beroende på hur djup processering man vill uppnå, alltså hur komplicerade relationer man vill hämta ur texten. För enkla ändamål räcker det med att man svarar på vissa förutbestämda frågor angående textens innehåll. För mer komplicerade ändamål krävs en mer allmän representation. En sådan kan vara första ordningens predikatlogik, där man på ett flexibelt sätt kan representera många olika slags relationer. Som ett enkelt exempel kan vi betrakta meningen “Apan äter en banan”. Uttryckt som en sats i predikatlogiken kunde denna händelse uttryckas på följande sätt:

Äter(Apan, Banan)

Satserna kan referera till varandra och således bildas en nätverksstruktur av informationen i texten. Genom att utvinna ett stort antal dylika satser får man en representation för det som sägs i hela texten [JM00]. Efter det kan man använda textgenereringsmetoder för att omvandla den formella representationen till en text i kortare utförande än den ursprungliga texten [SJ98].

I många fall räcker det dock med att bara svara på vissa förutbestämda frågor angående textens innehåll. Detta gäller speciellt om man på förhand vet hurdan sorts text som skall sammanfattas, och vilken typ av information som sammanfattningen skall innehålla. Då kan man använda en mall eller blankett med fält som fylls i på basen av texten som skall sammanfattas [JM00]. Allteftersom processeringen fortskrider, fyller man i fler och fler fält i blanketten. Fälten kan bekräftas av senare information i texten, eller någon ny information kan precisera eller kullkasta tidigare information. Slutprodukten är en färdigt ifylld blankett som kan användas som källmaterial för en textgenereringsalgoritm.

## 4.3 Lexikala kedjor, koherens och kohesion

Tidigare konstaterades att textutdrag kan leda till missvisande sammandrag där en mening som syftar på en tidigare mening får en felaktig betydelse om den står ensam. För SweSum [Dal00] finns en experimentell metod för att lösa detta problem: pronomenresolution. Metoden fungerar som en förhandsprocessor, och ersätter pronomen som han, hon, den, det, med de faktiska substantiv som de refererar till. Metoden bygger på identifikation av substantiv med hjälp av en speciell ordlista, och den använder ett minne ur vilket det rätta substantivet kan hämtas.

Metoden klarar dock bara av pronomenresolution för så kallade anaforiska relationer, det vill säga situationer där substantivet i texten kommer före alla de pronomen som syftar på det. Systemet känner igen motsatsen, kataforiska relationer, men eftersom den gör endast ett svep genom texten kan den inte utföra pronomenresolution på dem [Has00]. Kataforiska relationer hittas till exempel i meningar som denna: “Hon är rolig, din moster.” Här kan man inte veta vad pronomenet “hon” syftar på förrän man har läst hela meningen, men man kan känna igen ordet som ett pronomen och undvika att misstolka dess syftning.

Pronomenresolution är dock inte en fullständig lösning på syftningsproblemet. En mer generell metod, vars syfte är något annorlunda, är att hitta kedjor av betydelse.

Betydelsekedjor eller, som Jane Morris och Graeme Hirst i sin formella beskrivning kallar dem, lexikala kedjor (lexical chains) [MH91], bygger på det faktum att det i en text i regel finns en logisk följd av tankar som utvecklas allteftersom texten fortskrider. Man talar intuitivt om en "röd tråd" i texten, och en välskriven text kan ha flera sådana implicita "trådar" som en läsare kan uppfatta och som håller ihop resonemanget.

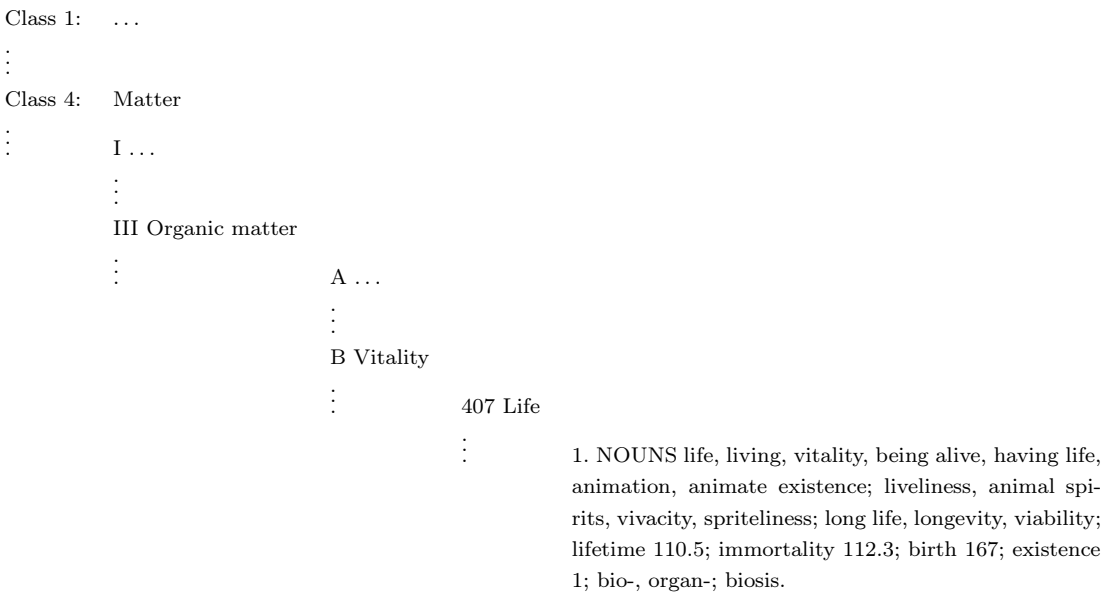
Man kan iaktta två egenskaper i en text som tillsammans bidrar till detta fenomen: koherens och kohesion. Kohesion innebär att olika element förekommer tillsammans. Koherens innebär att någonting är genomgående logiskt och har en vettig betydelse [MH91]. Meningarna i en välskriven text har en koherensrelation som bidrar till att läsaren uppfattar att de bygger på samma uppfattning och inte motsäger det som tidigare har sagts. Koherens och kohesion är drag i textens diskurs som gör att en allt mer detaljerad eller större bild av ämnet som behandlas byggs upp successivt. En text där koherensrelationerna och kohesionen är svag eller fattas, uppfattas som osammanhängande och svår att tolka.

Morris och Hirst menar att det inte finns en generell algoritmisk metod för att identifiera koherensrelationer, men att det är möjligt att identifiera en speciell sorts kohesion: lexikal kohesion. Lexikal kohesion uppstår via semantiska relationer mellan ord, alltså att orden på ett mätbart sätt hänvisar till samma betydelse. Lexikala kedjor är en sekvens av relaterade ord som kan ligga i samma mening, i närliggande meningar, eller spridda över hela texten och som uppvisar lexikal kohesion [MH91]. Detta påminner på konceptnivå om pronomenresolutionen i SweSum, men utsträcker sig till alla delar av texten. Syftet är inte direkt att undvika syftningsproblem, utan snarare att hitta en mer pålitlig poängsättningsmetod som plockar ut den "röda tråden" ur texten.

Följande exempel illustrerar en lexikal kedja:

> Apan äter en *banan*. *Bananen* är grön och lite rå. *Den* är ändå god, tycker apan. Alla *frukter* smakar inte alls lika gott.

I detta textfragment bildar orden "bananen", "den" och "frukter" en lexikal kedja som börjar vid den ursprungliga bananen i den första meningen. Det är alltså fråga om olika former av relationer till och syftningar på samma banan i alla meningarna. I den andra meningen är det fråga om en upprepning, i den tredje meningen en syftning med hjälp av pronomen, och i den fjärde en upprepning genom ett mer allmänt begrepp ("bananen är en sorts frukt").




Figur 7: Rogets tesaurus [Cha77] grupperar ord i kategorier. Den högsta nivån är indelad i åtta klasser. Klasserna indelas i underklasser (romerska siffror), som i sin tur ytterligare delas in i underklasser (bokstäver). Denna tredje nivå indelas i totalt 1042 numrerade baskategorier. Baskategorierna indelas i numrerade stycken med relaterade ord, och inom varje stycke kan mindre grupper separeras med semikolon. Förutom ord kan referenser till relaterade kategorier eller stycken förekomma. Figuren är adapterad från Morris och Hirsts beskrivning [MH91].

Morris och Hirst använder Rogets tesaurus [Cha77] för att identifiera lexikala kedjor. En tesaurus är en ordbok där orden grupperas enligt kategorier på olika nivåer. I figur 7 visas en liten del av tesaurusen, så som Morris och Hirst beskriver den. Till tesaurusen hör ett index med vilket man kan hämta ord som är relaterade till ett givet ord, de olika betydelserna för ordet och en referens till en kategori eller ett stycke för varje betydelse [MH91].

Huvuddragen i Morris och Hirsts metod är enkel och består av två steg som utförs för varje ord i texten i tur och ordning: Först kontrolleras om ordet är lämpligt för att tas med i en kedja. I valet tas prepositioner och mycket högfrekventa ord bort, och en morfologisk analys utförs så att olika böjningsformer av samma ord inte skall räknas som olika ord. Sedan används Rogets tesaurus för att kontrollera om ordet kan läggas till en existerande kedja, eller om ordet skall vara en kandidat för en ny kedja. Ord som inte på ett förutbestämt antal varv har bildat en kedja förkastas – alltså har inga relaterade ord hittats inom ett visst avstånd.

Rogets tesaurus används för att bestämma om ett ord skall läggas till en existerande

kedja. Ordet jämförs med det sista ordet i de redan existerande kedjorna. Fem olika relationer är möjliga och resulterar i att ordet kopplas till ett annat ord:

1. Orden har en gemensam kategori i sina index-noteringar.
2. Det ena ordet har en kategori i sin index-notering som i sin tur har en referens till en kategori som det andra ordet hör till.
3. Det ena ordet finns antingen i det andra ordets index-notering, eller i en av det andra ordets kategorier.
4. Orden är i samma grupp.
5. Orden har referenser till samma kategori i sina index-noteringar.

Morris och Hirst presenterade ingen konkret, automatiserad implementering av sin metod eftersom de inte hade tillgång till en maskinläsbar kopia av Rogets tesaurus [MH91], men senare forskning har använt metoden för praktiska försök. Ett problem med metoden är att den inte inte skiljer på olika betydelser av ord. Textstycken som använder samma ord i olika betydelser formar en enda kedja, då de istället borde forma olika kedjor. Valet av lexikal kedja innebär att ordet disambigueras implicit. Man kan lägga till ett explicit disambigueringssteg eller särskilt beakta mångtydiga ord då man använder tesaurusen, vilket förbättrar kedjornas kvalitet [BE97].

Lexikala kedjor kan alltså användas för att hitta den "röda tråden" i en text, och således ger de ett mått på meningarnas relevans. För att producera ett sammandrag väljer man de starkaste kedjorna och således de meningar som kedjans ord finns i. Det finns olika sätt att bestämma vilka kedjor som är starka. Längden på kedjan är ett enkelt mått, men mer komplicerade mått har också använts. I vissa implementeringar har WordNet använts istället för Rogets tesaurus [BE97, SM00, BCP01].

Att använda lexikala kedjor för att producera ett sammandrag är en metod som kombinerar aspekter hos både textutdrag och faktautvinning. Texten analyseras med utgångspunkt i språkets semantiska struktur, men själva sammanfattningen består av meningar ur den ursprungliga texten. Här knyts också disambiguering av ord och automatisk sammanfattning ihop – man kan inte producera en högklassig sammanfattning utan att förstå den ursprungliga texten.

# 5 Slutsatser

Språkprocessering är ett område av datavetenskapen där man för tillfället gör stora framsteg. I takt med att mängden språklig information i samhället ökar, ökar också behovet av att kunna processera språk effektivt. I denna avhandling har jag lyft fram två problem inom språkprocesseringen: disambiguering av ord och automatisk sammanfattning.

Disambiguering av ord och automatisk sammanfattning har många gemensamma kontaktytor. Båda problemområdena kan fungera på textens ytnivå eller på en semantisk nivå som analyserar textens betydelse djupare. Ofta kombineras båda metoderna i praktiska implementeringar. I automatisk sammanfattning kan disambiguering av ord bli nödvändigt för att välja ut rätt delar av texten.

Metoderna som används för både disambiguering av ord och automatisk sammanfattning liknar varandra. På ytnivån används statistiska metoder som på datorer ofta implementeras med matrisräkning. På den semantiska nivån används olika slag av relationer mellan ord eller begrepp, varav WordNet och olika slag av ontologier hör till de nuvarande populära informationskällorna.

Även om ett stort antal praktiska språkteknologiska applikationer finns, har relativt få nått slutanvändarnas datasystem. De flesta konsumentapplikationerna finns i bakgrunden av olika slags nyhets- och webbtjänster, och få program kan utföra avancerad språkprocessering. Bristen på tillgänglig programvara för till exempel automatisk sammanfattning av www-material, visar att språkprocesseringen ännu är i ett mycket tidigt stadium av sin utveckling. Även om vissa specifika applikationer för till exempel ordbehandling är i allmänt bruk, är ännu många av framtidsvisionerna för automatisk språkprocessering utom räckhåll. Det är också oklart hur långt avståndet till dessa framtidsvisioner verkligen är.

Vidare utveckling av området fortsätter mot ännu djupare analys av texten och kräver yttre kunskap som läggs till den information som finns i texten. Efter en viss gräns behövs allmän kunskap om den mänskliga världen och mänsklig kultur för att kunna processera språk på den önskade nivån. Å andra sidan finns det mycket som talar för fortsatt utveckling av ytlig processering, och inom många applikationsområden är en djup analys inte nödvändig för att nå ett användbart resultat. Metoderna kommer sannolikt att utvecklas sida vid sida och kombineras i olika former allteftersom förståelsen för det mänskliga språket ökar.

# Referenser


BCP01 Brunn, M., Chali, Y. och Pinchak, C., Text summarization using lexical chains. *Workshop on Text Summarization*, New Orleans, Louisiana, U.S.A., September 13–14 2001, URL `http://citeseer.ist.psu.edu/brunn01text.html`. Online proceedings available at http://www-nlpir.nist.gov/projects/duc/duc2001/agenda_duc2001.html.

BE97 Barzilay, R. och Elhadad, M., Using lexical chains for text summarization. *Proceedings of the Intelligent Scalable Text Summarization Workshop (ISTS'97)*, Madrid, Spain, 1997, URL `http://citeseer.ist.psu.edu/barzilay97using.html`.

CC98 Chen, J.-N. och Chang, J. S., A concept-based adaptive approach to word sense disambiguation. *Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics*, Boitet, C. och Whitelock, P., redaktörer, San Francisco, California, 1998, Morgan Kaufmann Publishers, sidorna 237–243, URL `http://citeseer.ist.psu.edu/chen98conceptbased.html`.

Cha77 Chapman, R. L., *Roget's International Thesaurus (Fourth Edition)*. Harper and Row, New York, 1977.

CR95 Church, K. W. och Rau, L. F., Commercial applications of natural language processing. *Communications of the ACM*, 38,11(1995), sidorna 71–79.

Dal00 Dalianis, H., SweSum – A Text Summarizer for Swedish. URL `http://www.dsv.su.se/~hercules/papers/Textsumsummary.html`. Nada, Kungliga Tekniska Högskolan, Stockholm, 2000.

DDL+90 Deerwester, S., Dumais, S., Landauer, T. K., Furnas, G. och Harshman, R., Indexing by Latent Semantic Analysis. *Journal of the American Society of Information Science*, 41,6(1990), sidorna 391–407. URL `http://citeseer.ist.psu.edu/deerwester90indexing.html`.

Faw92 Fawcett, R. P., Review of "A theory of computer semiotics: semiotic approaches to construction and assessment of computer systems" by P.

B. Andersen. Cambridge University Press 1990. *Computational Linguistics*, 18,4(1992), sidorna 555–562. Reviewer-Robin P. Fawcett.

Fel98 Fellbaum, C., *WordNet, an Electronic Lexical Database*. MIT Press, 1998.

FK64 Francis, W. N. och Kucera, H., *A standard sample of present-day English for use with digital computers*. Brown University, 1964.

Has00 Hassel, M., Pronominal Resolution in Automatic Text Summarization. Pro gradu, Department of Computer and Systems Sciences, Stockholm University, 2000. URL `http://www.nada.kth.se/~xmartin/papers/Master-PRM.PDF`.

HC81 Hendrix, G. G. och Carbonell, J. G., A tutorial on natural-language processing. *ACM 81: Proceedings of the ACM '81 conference*. ACM Press, 1981, sidorna 4–8.

IV98 Ide, N. och Véronis, J., Introduction to the special issue on word sense disambiguation: the state of the art. *Computational Linguistics*, 24,1(1998), sidorna 2–40.

JM00 Jurafsky, D. S. och Martin, J. H., *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Prentice-Hall, Inc., Upper Saddle River, New Jersey, 2000.

MH91 Morris, J. och Hirst, G., Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17,1(1991), sidorna 21–48.

SJ98 Sparck-Jones, K. *Advances in Automatic Text Summarization*, kapitel Automatic summarising: factors and directions. MIT press, 1998. URL `http://citeseer.ist.psu.edu/jones98automatic.html`.

SM00 Silber, H. G. och McCoy, K. F., Efficient text summarization using lexical chains. *Intelligent User Interfaces*, 2000, sidorna 252–255, URL `http://citeseer.ist.psu.edu/silber00efficient.html`.

Swi97 Swift, J., *Gulliver's Travels into Several Remote Nations of the World*. Project Gutenberg, Oxford, Mississippi, 10:e utgåva, 1997. URL

http://www.gutenberg.org/dirs/etext97/gltrv10h.htm. Transcribed from the 1892 George Bell and Sons edition by David Price.

Zec95 Zechner, K., Automatic Text Abstracting by Selecting Relevant Passages. Pro gradu, Center for Cognitive Science, University of Edinburgh, 1995. URL http://www-2.cs.cmu.edu/~zechner/abstr.pdf.

# A SweSum: exempel på automatiskt sammandrag

Följande automatiska sammandrag har producerats av SweSum – Automatisk Textsammanfattare av Martin Hassel och Hercules Dalianis (http://swesum.nada.kth.se/) [Dal00]. Systemet ombads sammanfatta texten med en sammanfattningsgrad av 13%. Inga nyckelord gavs, utan SweSum har själv identifierat nyckelorden med hjälp av sin sammanfattningsalgoritm. Tyvärr kunde inte pronomenresolution användas på grund av att funktionen inte var tillgänglig. I gengäld demonstreras problemet med bristfällig koherens med ett exempel som inte kunde ha åtgärdats med pronomenresolution.

Den ursprungliga texten är från Hufvudstadsbladets ledarsida 14.3.2005. (Från Hbl:s webbplats http://www.hbl.fi/ 14.3.2005.)

## 1 Ursprunglig text

Meningen är att bildligt kasta hela välfärdssamhället upp i luften.

**Kommunens roll måste skrivas om**

Regeringen vill reformera kommun- och servicestrukturerna i landet. Bakom den meningen i torsdagens beslut om budgetramen för 2006 döljer sig en omfattande process som ingen ännu kan se slutet på.

Det officiella startskottet har nu gått för en djuplodande omprövning av kommunernas roll. Officiellt sägs syftet vara att säkra välfärdstjänsterna i hela landet. Men syftet kunde lika väl sägas vara att krympa den offentliga sektorn – eftersom pengarna inte räcker till.

Samtidigt har folket just upplysts om att exporten drar bra och att finansministeriet har skrivit upp sin tillväxtprognos. Så har den förestående strukturomvandlingen inte heller så mycket att göra med tillgången på pengar i landet.

Problemet är snarare demografiskt. Det handlar om landsbygdens avfolkning och om de stora årskullarnas förestående pensionering.

Det har inte rått brist på antydningar om det som komma skall. Bland annat slog regeringens utkast till redogörelse över förvaltningens problem redan i januari fast att kommunsektorn inte har effektiverat sin verksamhet tillräckligt snabbt.

I det fortsatta arbetet på den redogörelsen föddes tanken på ett projekt, snarare än den parlamentariska utredning som Finlands kommunförbund ville ha. Tanken är nu att statssekreterarna vid de berörda ministerierna före utgången av april kommer med ett förslag om projektets ledning och exakta uppdrag. Man fö-

reställer sig en styrgrupp med bred politisk förankring.

I första skedet skall all tillgänglig kunskap sammanställas. Social- och hälsovårdens forskningscentral Stakes har till exempel redan kört i gång ett eget projekt för att staka ut välfärdsstatens gränser. Därefter följer ett planeringsskede då kommunernas representanter på regional nivå skissar upp nya mönster för verksamheten.

Meningen är – varken mer eller mindre – att bildlikt kasta hela välfärdssamhället upp i luften. När det kommer ner igen får man se hur bitarna passar ihop.

Statsledningen hoppas att själva processen skall generera insikt i problemets natur snarare än protester och det så vanliga motståndet mot förändringar. Detta ställer exceptionellt höga krav på öppenhet och delaktighet.

Finland är ingen pionjär i fråga om att ompröva den nordiska välfärdens strukturer. I Sverige tillsattes en parlamentarisk utredning, den så kallade Ansvarskommitten, redan för flera år sedan. Ett första delbetänkande avgavs 2003 och arbetet fortsätter med att bland annat definiera vad kommunen skall göra och hur arbetet skall fördelas mellan staten, den regionala nivån och kommunerna.

Danmark har en radikal kommunreform på gång och i Norge har motsvarande behov konstaterats, också om den politiska viljan låter vänta på sig.

Till saken hör att Sverige, Norge och Danmark alla genomförde stora kommunreformer redan under välfärdssamhällets tillblivelse på 1960- och 1970–talet. Det samma planerades i Finland, men gick på många håll inte att genomdriva. Det var då centern – dåvarande agrarförbundet – fick sitt smeknamn "kommunpartiet".

Visserligen reducerades antalet kommuner kraftigt i både svenska Österbotten och Västnyland. Men bland annat i Åboland är strukturen lika illa splittrad som någonsin i det centerdominerade inlandet.

Nu blåser andra vindar, kyligare vindar. Region- och kommunminister Hannes Manninen (c) talar med nästan samma tunga som finansminister Antti Kalliomäki (sdp). Manninen vill klarlägga kommunernas basuppgifter och hur de bäst kan produceras så att den galopperande utgiftsökningen kan tyglas. Kalliomäki kräver stabilitet och förutsägbarhet i de ekonomiska relationerna mellan stat och kommun.

En ny statsandelsreform skulle träda i kraft 2006. Den har senarelagts, med motiveringen att ingen vettig reform kan genomföras om man inte kan anslå mera pengar för finansiering av kommunernas verksamhet. I offentligheten har detta tolkats som att det kanske senare går att få fram de behövliga tilläggsanslagen.

Det är att vara överoptimistisk.

Den rätta tolkningen är att regeringen först vill se hur mycket den offentliga sektorn kan spara, genom att omdefiniera sitt uppdrag och genom att skapa större befolkningsunderlag för förvaltningsenheterna. Först därefter är det dags att ändra på finansieringssystemets detaljer.

JOHANNA WESTMAN

## 2 Automatiskt sammandrag

Meningen är att bildligt kasta hela välfärdssamhället upp i luften.

Kommunens roll måste skrivas om

Regeringen vill reformera kommun- och servicestrukturerna i landet. Ett första delbetänkande avgavs 2003 och arbetet fortsätter med att bland annat definiera vad kommunen skall göra och hur arbetet skall fördelas mellan staten, den regionala nivån och kommunerna.

Den rätta tolkningen är att regeringen först vill se hur mycket den offentliga sektorn kan spara, genom att omdefiniera sitt uppdrag och genom att skapa större befolkningsunderlag för förvaltningsenheterna.

JOHANNA WESTMAN

Lexikon: Svenska
Ord före 664
Ord efter 88
Sammanfattningsgrad: 13%
Typ av text: tidningstext
Nyckelord: kommun skall arbete offentlig problem projekt regering först mening mycken

## 3 Kort analys

Om man jämför meningen “Ett första delbetänkande avgavs...” i sammandraget med samma mening i den ursprungliga texten, märker man att meningen i sammandraget verkar syfta på den finska regeringen, kommunen och staten, medan den i själva verket i originaltexten visar hur situationen ser ut i Sverige.

Även om pronomenresolution hade använts, så hade inte problemet kunnat åtgärdas, eftersom syftningen på den tidigare meningen är helt implicit.