Disambiguierung deutschsprachiger Diskursmarker: Eine Pilot-Studie

Petra Saskia Bayerl (Gießen)

Abstract

Discourse markers such as German *aber, wohl* or *obwohl* can be regarded as valuable information for a wide range of text-linguistic applications, since they provide important cues for the interpretation of texts or text segments. Unfortunately, many of them are highly ambiguous. Thus, for their use in applications like automatic text summarizations a reliable disambiguation of discourse markers is needed. This should be done automatically, since manual disambiguation is feasible only for small amounts of data.

The aim of this pilot study, therefore, was to investigate methodological requirements of automatic disambiguation of German discourse markers. Two different methods known from word-sense disambiguation, Naive-Bayes and decisionlists, were used for the highly ambiguous marker *wenn*. A statistical approach was taken to compare the two approaches and different feature combinations.

1 Einleitung

Diskursmarker wie *wenn*, *sogar* oder *aber* können als Signale angesehen werden, welche Hinweise geben auf die funktionale Beziehung bzw. rhetorische Relation, die zwischen zwei Textelementen besteht, und geben somit Hilfestellungen für die Interpretation von Aussagen oder Propositionen (Millis/Golding/Barker 1995). Die Art solcher funktionaler Beziehungen zwischen Textelementen läßt sich in einfacher Weise beschreiben über das Instrumentarium der *Rhetorical Structure Theory* (RST) nach Mann und Thompson (1988). Mit Hilfe der dort formulierten Relationen und Formalismen ist es möglich, die rhetorische Struktur von Texten baumartig als Abfolge zusammenhängender Textteile abzubilden (s. Abbildung 1).

ISSN 1615-3014

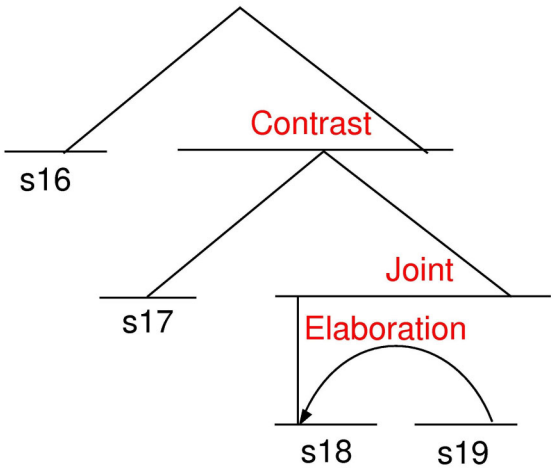

Abbildung 1: Rhetorische Struktur eines Teiltextes mittels RST

Da solche Relationen häufig über Diskursmarker vermittelt werden, bieten diese wertvolle, konkrete Ansatzpunkte für die Analyse von Diskursstrukturen und können damit als eine wertvolle Informationsquelle in zahlreichen Bereichen der Texttechnologie angesehen werden. So wird etwa im Kontext der automatischen Textzusammenfassung versucht, durch die automatische Erkennung von Diskursmarkern und die Zuweisung einer einzelnen rhetorischen Relation, die Bedeutung von Textteilen im Gesamttext zu bestimmen, um so sinnvolle Kurzfassungen von Texten zu erstellen (z. B. Marcu 2000). Auch die einfache automatische Analyse von Strukturen (Kurohashi/Nagao 1994; Mann/Matthiessen/Thompson 1989) oder die automatische Generierung von Texten (Hovy 1993) greifen auf ähnliche Mechanismen zurück.

Als problematisch erweist sich allerdings, daß Diskursmarker sich häufig nicht eindeutig auf eine einzelne rhetorische Relation zurückführen lassen, sondern der gleiche Marker vielmehr in Abhängigkeit vom Kontext unterschiedliche Relationen abbildet. So signalisiert etwa wenn in Satz 1 des folgenden Beispiels gemäß RST eine kausale (*volitional cause*) Beziehung, in Satz 2 hingegen eine einschränkende (*restriction*) Beziehung.

(1) Präsident Heinz Weisener drohte, den Trainer zu feuern, wenn sein Team nicht vier Punkte aus den nächsten beiden Partien holen würde.

(2) Die BRD wird ihren Titel nicht verteidigen können, zumindest wenn es nach dem Gros unserer Befragten geht.

Ähnliches läßt sich bei den meisten Diskursmarkern (z.B. *da*, *weil*, *aber*) finden. Diese Ambiguität in der Funktion von Diskursmarkern kann bei texttechnologischen Anwendungen zu Problemen führen, da die Erkennung der Relationen keine eindeutigen Ergebnisse mehr ergibt und sich dadurch Ungenauigkeiten oder gar Fehler bei der Repräsentation von Textstrukturen ergeben können.

Es stellt sich folglich die Frage, ob es möglich ist, solche Ambiguitäten mit ausreichender Genauigkeit aufzulösen, um die Informationen, die durch Diskursmarker im Text repräsentiert werden, besser nutzen zu können. Ziel wäre somit die automatische Analyse von Texten oder Korpora und die Zuweisung eindeutiger Relationen an erkannte Diskursmarker. Erste Ansätze hierfür könnten Methoden liefern, wie sie für die Disambiguierung von Polysemen und

Homonymen entwickelt wurden. Auch hier geht es um die Auswahl der 'korrekten' Bedeutung aus einer ganzen Reihe möglicher Verwendungsweisen eines Wortes (z. B. *Schule* als Bildungsinstitut vs. Schwarm von Jungfischen). Die unterschiedlichen rhetorischen Funktionen eines Diskursmarkers ließen sich hierbei in etwa mit den unterschiedlichen Verwendungsweisen eines Wortes vergleichen, so daß es durchaus nahe liegt, bisherige Methoden der Wortdisambiguierung auf diesen Fall zu übertragen. Weiterhin könnten sich durch die Analyse solcher Methoden erste Hinweise auf die Gestaltung von Korpora ergeben, die im Rahmen texttechnologischer Anwendungen eingesetzt werden sollen.

2 Theoretische Vorüberlegungen

2.1 Verwendungsspektren

Der erste Schritt im Rahmen einer Disambiguierung besteht in der Bestimmung des Sets unterschiedlicher Verwendungsweisen, die zu einem Wort gehören. Hierfür werden häufig papierene oder maschinenlesbare Lexika wie *WordNet* (Miller 1990; Fellbaum 1998) oder Thesauri herangezogen, die ein Spektrum an Verwendungsweisen mit Definitionen der verschiedenen Einzelbedeutungen verbinden. Neben der Anzahl unterschiedlicher Verwendungsweisen können hier auch Anhaltspunkte für die Differenzierung der Verwendungsweisen extrahiert werden.

Vergleichbare Lexika für Diskursmarker, die eine Auflistung möglicher rhetorischer Relationen und Verwendungskontexte liefern, existieren bislang nur im Anfangsstadium (Stede/Umbach 1998). Elektronische Lexika im Sinne von WordNet liegen m.W. noch nicht vor. Für eine Untersuchung wie die vorliegende, müssen mögliche Spektren von rhetorischen Relationen für einzelne Diskursmarker deshalb zunächst aus einem gewählten Korpus extrahiert werden. Vollständigkeit kann in diesem Stadium deshalb bestenfalls angestrebt, aber nicht gewährleistet werden.

2.2 Korpora

Automatische Disambiguierung ist im allgemeinen angewiesen auf Korpora von einigen zehn- bis hunderttausend Trainingsbeispielen, die, meist in mühevoller Handarbeit, mit Verwendungsweisen für einzelne Wörter annotiert wurden. Während diese für Polysemdisambiguierungen als Auszüge aus dem *Brown*-Korpus oder dem *Wall Street Journal* zumindest für den englischsprachigen Raum vorliegen (vgl. Ide/Veronis 1998), existiert derzeit für (deutschsprachige) Diskursmarker nichts Vergleichbares. Die Datengrundlage, auf die sich eine Untersuchung zur Disambiguierung von Diskursmarkern stützt, kann deshalb nur aus selbstannotierten Beispielen bestehen, wobei die endgültige Größe des so erstellten Korpus abhängig ist von der verfügbaren Zeit und den bereit stehenden Ressourcen.[1]

[1] Ng (1997) errechnete etwa 16 Mannjahre für die Annotation einer ausreichend großen Datenmenge, was er jedoch als einen der Sache durchaus angemessenen Aufwand beschreibt.

2.3 Disambiguierungsverfahren

Unter Disambiguierungsverfahren sind hier jene Methoden zu verstehen, welche in der Lage sind, einem ambigenWort die (mit einer gewissen Genauigkeit) wahrscheinlichste Bedeutung automatisch zuzuweisen und es damit semantisch eindeutig zu bestimmen. Disambiguierungsverfahren greifen hierzu im allgemeinen auf zwei Informationsquellen zurück (s.a. Ide/Veronis 1998): zum einen werden Informationen aus dem Kontext des ambigen Wortes betrachtet. Hierzu gehören zum Beispiel Stellungsinformationen (etwa Position am Anfang oder Ende eines Satzes), umgebende Wörter (Kollokationen) oder morpho-syntaktische Informationen. Zum anderen werden häufig externe Quellen herangezogen, hier vor allem Lexika, Enzyklopädien oder auch Übersetzungen. Der erste Ansatz wird im allgemeinen als daten- oder korpus-basiert bezeichnet, während letzterer als wissensbasiert gilt (vgl. Wilks/Stevenson 1997). In zahlreichen Studien werden beide Verfahren auch kombiniert, um größere Effektivität zu erreichen (z.B. Li/Szpakowicz/Matwin 1995; Wilks/Stevenson 1998).

Da bislang geeignete externe Quellen für Diskursmarker nicht in ausreichendem Maße zur Verfügung stehen, kann für den Zweck der Diskursmarkerdisambiguierung nur auf daten-basierte Verfahren zurückgegriffen werden. Diesen gemeinsam ist die Überlegung, daß sich unterschiedliche Verwendungsweisen oder Bedeutungen eines Wortes in unterschiedlichen Kontexten manifestieren. So ist etwa die Wahrscheinlichkeit, mit der Wörter wie *Wasser* oder *Fluß* im Kontext von *Bank* im Sinne 'Flußufer' auftreten, größer als bei *Bank* im Sinne von 'Geldinstitut'. Solche unterschiedlichen Wahrscheinlichkeiten lassen sich für verschiedene Merkmale aus bekannten, d.h. bereits disambiguierten Beispielen extrahieren und so mittels statistischen Verfahren auf bisher unbekannte Beispiele übertragen.

Unterschiede zwischen den einzelnen statistisch orientierten Verfahren bestehen im wesentlichen in der Art der betrachteten Merkmale und ihrer Gewichtung bei der Entscheidung über die jeweilige Verwendungsweise eines bislang unbekannten ambigen Wortes (einen guten Überblick über Disambiguierungsmethoden bieten z.B. Ide/Veronis 1998; Manning/Schütze 1999). Die Güte der Verfahren ist im wesentlichen abhängig von der Art der verwendeten statistischen Methoden, der Größe des Trainingssamples und evtl. verfügbaren externen Informationsquellen.

Da sich die Grundcharakteristiken der beiden Aufgaben Disambiguierung von Polysemen/Homonymen und Disambiguierung von Diskursmarkern nicht wesentlich unterscheiden, dürften die mathematischen Überlegungen, die auf dem Gebiet polysemer/homonymer Wörter angestellt wurden, auch auf das der Diskursmarker übertragbar sein. Da im Bereich der Diskursmarkerdisambiguierung jedoch bislang kaum Erfahrungen hierzu vorliegen, muß zunächst geklärt werden, welche Methoden für die Diskursmarkerdisambiguierung geeignet sind und welche Merkmale bzw. Merkmalskombinationen am meisten zur Disambiguierung beitragen können. Auf diese Weise ließe sich auch entscheiden, welche Informationen in Korpora zur Diskursmarkerdisambiguierung eingefügt oder aber von extern bezogen werden müßten, um zu optimalen Ergebnissen zu kommen.

ISSN 1615-3014

3 Methodik

3.1 Korpus

Grundlage für die vorliegende Untersuchung bildete das Deutsche Referenzkorpus (Dereko), welches Zeitungstexte der TAZ aus dem ersten Halbjahr 1994 enthält (Dipper et al. 2002). Die einzelnen Texte liegen in XML-Format vor und wurden nach einem an der Universität Tübingen entwickelten Textauszeichnungssystem annotiert (Ule 2002), welches eine Erweiterung des Corpus Encoding Standards (CES) (Ide/Priest-Dorman 2000) darstellt. Die Annotation selbst lief automatisch ab. Das Korpus stellt neben einer Grobeinteilung in Abschnitte und Sätze auch morphosyntaktische Informationen auf unterschiedlichen Ebenen zur Verfügung (s. Beispiel 1). So wird etwa jedes Wort als token mit seiner Wortformen, d. h. einem sog. *part-of-speech*-Tag (pos-Tag) und seiner Grundform (baseform) versehen. Des weiteren werden morphologische Informationen (morph) annotiert. Eine Besonderheit der Auszeichnung in Dereko stellen die sog. chunks dar, welche "non-recoursive continuous kernels of phrases" zusammenfassen und mit einem entsprechenden Label versehen.[2]

```
<ch c='VCLAF'>
  <token form='ist' info='' tnum='7212906'>
    <pos cert='1' rank='1' tag='VAFIN'>
     <baseform form='sein'>
       <morph desc='3s'>
       </morph>
      </baseform>
    </pos>
  </token>
</ch>
```

Beispiel 1: Ausschnitt aus dem Dereko-Korpus (leicht überarbeitet und gekürzt)

3.2 Material

Zur Disambiguierung ausgewählt wurde der Diskursmarker wenn, da er eine vielfältige Struktur markierter Relationen aufweist, die eine präzise Disambiguierung erfordert. Als Arbeitsgrundlage dienten 212 Sätze, die diesen Diskursmarker enthalten. Die Sätze wurden durch Zufallsauswahl aus dem Dereko-Korpus extrahiert, wobei neben dem eigentlichen Diskursmarkersatz auch jeweils ein Satz direkt vor und direkt nach dem Diskursmarkersatz entnommen wurde, um durch den Kontext die Zuordnung einer rhetorischen Relation zu erleichtern. Da einige der Vor- und Nachsätze ebenfalls ein wenn enthielten, lagen mit den 212 Satztripeln insgesamt 250 Diskursmarkersätze vor.

3.3 Disambiguierungsmethoden

Verwendet wurden die Algorithmen 'Naive-Bayes' und 'Decisionlists', die beide zu den überwachten (supervised) Methoden gehören.

[2] Für eine umfassende Beschreibung des Annotationssystems sei auf Ule (2002) oder Müller (2002) verwiesen.

3.3.1 Naive-Bayes

Das Naive-Bayes-Verfahren geht von der Annahme aus, daß alle Merkmale im Kontext eines Diskursmarkers einen eigenen, unabhängigen Beitrag zur Unterscheidung von Verwendungsweisen bzw. Relationen liefern (kritisch zur Annahme der Unabhängigkeit s. Domingos/Pazzani 1996). Der individuelle Beitrag eines Wortes errechnet sich über Maximum-Likelihood-Schätzungen wie folgt:

$$P(Merkmal_k|Relation_i) = \frac{C(Merkmal_k, Relation_i)}{C(Relation_i)}$$

als Auftretenswahrscheinlichkeit *P* eines Merkmals *k* im Kontext einer Relation *i* innerhalb des Trainingssamples mit *C* als Auftretenshäufigkeit im Trainingssample. Die Wahrscheinlichkeit, daß ein Diskursmarker im gegebenen Kontext *c* eine bestimmte Relation *i* markiert, wird bestimmt über:

$$\begin{aligned} P(Relation_i, Kontext_c) &= log\ P(Relation_i) \\ &\quad + \Sigma_{(Merkmal_k\ in\ c)}\ log\ P(Merkmal_k|Relation_i) \end{aligned}$$

mit

$$P(Relation_i) = \frac{C(Relation_i)}{C(Diskursmarker)}$$

Im Rahmen der Disambiguierung wird diejenige Relation ausgewählt, deren Wahrscheinlichkeit bei gegebenem Kontext maximal ist, wodurch die Irrtumswahrscheinlichkeit minimal gehalten werden kann.

3.3.2 Decisionlists

Die Disambiguierung mit dieser Methode basiert auf der Auswahl desjenigen Merkmals, welches die meiste Information über das betrachtete Wort enthält. Der Informationsgehalt eines Merkmals k wird als Gewicht w bezeichnet und errechnet sich nach einer von Yarowsky (1994) entwickelten Formel, die von Agirre und Martinez (2000) für mehr als zwei Verwendungsweisen adaptiert wurde, wie folgt:

$$w(Relation_i, Merkmal_k) = log\ \left(\frac{P(Relation_i|Merkmal_k)}{\Sigma\, P(Relation_{\neg i}|Merkmal_k)}\right)$$

Die Auflistung der Gewichte $w(Relation_i,Merkmal_k)$ ergibt die sog. Decisionlist. Das Merkmal mit dem höchsten Gewicht im Kontext eines Diskursmarkers bestimmt die zuzuweisende Relation. Im Gegensatz zur Naive-Bayes-Methode beruht die Entscheidung also lediglich auf einem einzelnen Merkmal, das die meiste Information beinhaltet.

Der Decisionlists-Ansatz hat sich in dem von Kilgarriff großangelegten Methodenvergleich Senseval II als eine der erfolgreichsten erwiesen (s. http://www.sle.sharp.co.uk/senseval2/). Dagegen ist Naive-Bayes eine, wie Mooney (1996) und Escudero, Marquez und Rigau (2000) zeigen konnten, äußerst einfache, aber dennoch sehr effektive Disambiguierungsmethode. Der Vergleich dieser beiden Methoden sollte vor allem zeigen, wie viele Merkmale bzw. Informationen zur Entscheidung über die Art der rhetorischen Relation notwendig sind.

3.4 Merkmalssets

Die Annotation des Dereko-Korpus ermöglicht den Zugriff auf Informationen wie Satzgrenzen, morphologische Informationen, pos-Tags und sog. chunks (ch), die Kernelemente einer Phrase repräsentieren (vgl. Ule 2002). Basierend auf den vorhanden Informationen wurden sieben verschiedene Merkmalssets gebildet, welche sich aus Variationen von Werten in den Bereichen Merkmale, Position und Kontextgröße bzw. Distanz zusammensetzen.

3.4.1 Merkmale

Um Größe bzw. Abstraktionsgrad der betrachteten Einheiten im Kontext eines Diskursmarkers zu variieren, wurden neben den Worten eines Satzes (sog. `token`) auch `pos`- und `ch`-Tags betrachtet. Neben der individuellen Betrachtung einzelner Merkmale wurden auch alle Kombination dieser drei Merkmale untersucht (`token+pos`, `token+ch`, `pos+ch`, etc). Gemäß Arbeiten von Marcu (1998, 2000) können das orthographische Umfeld und die Stellung des Diskursmarkers im Satz ebenfalls wichtige Hinweise auf seine Verwendungsweise liefern, weshalb diese beiden Elemente bei allen Merkmalssets in die Betrachtung miteinbezogen wurden.

3.4.2 Position

Neben der Art der umgebenden Einheiten kann auch deren Stellung im Satz Informationen über die aktuelle Diskursrelation liefern. Es stellt sich also die Frage, in welchem Maße die Position eines Merkmals in Relation zum Diskursmarker zur Unterscheidbarkeit beiträgt. Aus diesem Grund wurde die Position in verschiedenem Ausmaß in die Untersuchung miteinbezogen. Zum einen wurde nur das Auftreten von Worten im Kontext des Diskursmarkers betrachtet, während ihre Position keine Rolle spielte (sog. *bags*). Eine zweite Stufe unterschied lediglich zwischen der Stellung vor und nach dem Diskursmarker, während die dritte eine Differenzierung nach der genauen Position vor und nach dem Diskursmarker zuläßt. Diese Stufen werden im Folgenden als *bag*, *grob* und *fein* bezeichnet.

3.4.3 Distanz

Arbeiten von Yarowsky (1992, 1994) weisen darauf hin, daß je nach Art der Ambiguität ät Kontexte sehr unterschiedlichen Ausmaßes betrachtet werden müssen, um gute Ergebnisse zu erhalten. Während lokale Ambiguitäten mit kleinen Werten von drei oder vier Merkmalen vor und nach dem Diskursmarker auskommen, bedürfen semantische Ambiguitäten Fenstergrößen bis zu 50 Wörtern vor und nach dem ambigen Wort.

Da bisher im Bereich der Disambiguierung deutschsprachiger Diskursmarkern keine vergleichbaren Untersuchungen vorliegen, wurde mit mehreren Werten experimentiert. Zum einen bezog sich das Kontextfenster entweder nur auf den Diskursmarkersatz selbst, unabhängig von dessen Länge (Distanz *klein*), oder auf das gesamte Satztripel (Distanz *groß*). Des weiteren wurden folgende Fenstergrößen (Distanz *Fenster*) gewählt: 2, 3, 4, 5, 6, 8, 12, 20, 30, 40, 50.

Da davon auszugehen ist, daß die drei Bereiche Merkmale, Position und Distanz nicht unabhängig voneinander sind, wurden Kombinationen verschiedener Werte getestet, um zum einen den jeweiligen Beitrag eines Bereichs für die Disambiguierung zu untersuchen und zum anderen die bestmöglichen Kombinationen zu ermitteln. Aufgrund der Tatsache, daß im Rahmen der Decisionlists ein Merkmal sowohl durch Art als auch seine Position definiert ist (Yarowsky 1999), wurde hier lediglich mit der Positionsstufe fein gearbeitet. Eine Übersicht über die verwendeten Merkmalssets bei beiden Methoden bietet Tabelle 1.

Naive-Bayes

Distanz	*Position*	*Merkmal*
klein	fein	token, pos, ch
	grob	token+pos, token+ch
	bag	pos+ch, token+pos+c
groß	fein	token, pos, ch
	grob	token+pos, token+ch
	bag	pos+ch, token+pos+c
Fenster	fein	token, pos, ch
	grob	token+pos, token+ch
	bag	pos+ch, token+pos+c

Decisionlists

Distanz	*Position*	*Merkmal*
klein	fein	token, pos, ch
		token+pos, token+ch
		pos+ch, token+pos+c
groß	fein	token, pos, ch
		token+pos, token+ch
		pos+ch, token+pos+c
Fenster	fein	token, pos, ch
		token+pos, token+ch
		pos+ch, token+pos+c

Tabelle 1: Merkmalssets nach Methoden

3.4.4 Smoothing

Liegen nur wenige Trainingsdaten vor, spiegeln die Werte, die sich mittels Likelihood-Ratio errechnen lassen, selten die realen Verhältnisse der Sprache wieder. Mit hoher Wahrscheinlichkeit sind seltene Ereignisse aufgrund der geringen Datengrundlage gar nicht erst im Trainingssample enthalten. Um trotzdem reliable Ergebnisse zu erhalten, kann Ereignissen, die im Test-, nicht aber im Trainingskorpus auftreten, eine geringe

Restwahrscheinlichkeit zugewiesen werden. Diese Methode wird als discounting oder auch smoothing bezeichnet.

Die hier gewählte Methode ist das von Good (1953) entwickelte Good-Turing-Verfahren. Es schätzt die wahre Häufigkeit von n-Grammen in der Sprache über die Häufigkeit von n-Grammen im Trainingskorpus (für eine Einführung siehe z.B. Manning/Schütze 1999). In der vorliegenden Studie wurden lediglich Bigramme berechnet, größere Einheiten wurden nicht betrachtet.

3.5 Durchführung

3.5.1 Erstellung von Trainings- und Testkorpus

Die 212 Satzbeispiele wurden auf zwei Korpora aufgeteilt. 142 Satztripeln dienten als Trainingskorpus, während die verbleibenden 70 Beispiele als Testkorpus verwendet wurden. Da einige Vor- und Nachsätze ebenfalls den Diskursmarker wenn enthielten, erhöhte sich die Zahl der Diskursmarkersätze auf 169 bzw. 81, so daß insgesamt 250 Diskursmarkersätze vorlagen. Das Trainingskorpus wurde per Hand mit den durch wenn markierten rhetorischen Relationen annotiert, wofür ein neues Attribut rrel in das token-Element eingefügt wurde. Das Testsample selbst blieb unannotiert. Allerdings wurden eine Kopie hiervon in gleicher Weise wie das Trainingskorpus annotiert, um die Richtigkeit der im Rahmen der Disambiguierung zugewiesenen Relationen überprüfen zu können. Im Ganzen wurden elf rhetorische Relationen zugewiesen. (Eine Liste der verwendeten Relationen findet sich in Anhang A.)

3.5.2 Disambiguierung

Die eigentliche Disambiguierung erfolgte mittels Perl-Programmen für die beiden Methoden und die unterschiedlichen Merkmalssets. Für jede Merkmalskombination wurden, basierend auf dem Trainingssample, die mittels Good-Turing angepaßten Wahrscheinlichkeiten berechnet. Auf Grundlage dieser Werte durchlief jede Methode für jedes seiner Merkmalssets einmal das Trainingskorpus, um die Trainingsparameter zu bestimmen. Im zweiten Durchlauf wurde das Testkorpus eingelesen und für jeden auftretenden Diskursmarker gemäß der in 3.3.1 und 3.3.2 dargestellten Kennwerte die wahrscheinlichste rhetorische Relation für dieses Beispiel ausgewählt. Die gewählte Relation wurde zusammen mit der token-Nummer des entsprechenden wenn-Elements gespeichert, um eine eindeutige Zuordnung zu gewährleisten. Um überprüfen zu können, ob das gewählte Smoothing-Verfahren sich als effektiv erweist, wurden alle Disambiguierungen ein weiteres Mal ohne Smoothing durchgeführt.

3.6 Evaluation

Um die Güte des jeweiligen Algorithmus zu bestimmen, wurden mit Hilfe der annotierten Form des Testsamples für jedes Diskursmarkervorkommen diejenige rhetorische Relation bestimmt, die für das jeweilige Beispiel als 'richtig' anzusehen ist. Anschließend wurden die durch den Algorithmus gewählten Relationen mit den korrekten Werten verglichen. Als einfachste Form der Gütebestimmung bietet sich der prozentuale Anteil korrekter

Zuweisungen an. Dieser ermöglicht einen raschen, leicht interpretierbaren Vergleich sowohl über verschiedene Methoden hinweg als auch mit der sog. Baseline. Diese errechnet sich als

$$Baseline = \frac{C(\text{häufigste Relation des DM})}{C(\text{alle DM im Testsample})}$$

mit DM - Diskursmarker und beschreibt den Anteil korrekt annotierter Relationen, der durch Zuweisung der am häufigsten auftretenden Relation an alle Testbeispiele erreicht würde, d.h. ohne daß eine Disambiguierung vorgenommen würde. Für die kleine Fenstergröße betrug die Baseline 51,3 %, für das Satztripel 50,0 %.

4 Ergebnisse

Insgesamt wurden 728 Disambiguierungen durchgeführt. Hierbei entfielen 546 auf die Naive-Bayes-Methode, 182 auf Decisionlists. Jeweils die Hälfte der Disambiguierungen erfolgten mit bzw. ohne Smoothing. Beide Methoden erreichten mit allen verwendeten Merkmalssets bestenfalls die von der Baseline vorgegeben 51,3% bzw. 50,0% korrekten Zuweisungen (s. Tabelle 2). Naive-Bayes erreichte im Mittel 34,5% (s = 21,6) korrekte Zuweisungen, Decisionlists 9,6% (s = 13,5).

Algorithmus	N	Mittel	Standard-abweichung	Minimum	Maximum
Naive-Bayes					
insgesamt	546	34,5	21,6	0,0	51,3
ohne Smoothing	273	27,7	23,5	0,0	51,3
mit Smoothing	273	41,4	16,9	0,0	51,3
Decisionlists					
insgesamt	182	9,6	13,5	0,0	51,3
ohne Smoothing	91	18,2	14,5	2,9	51,3
mit Smoothing	91	0,9	2,3	0,0	10,0

Tabelle 2: Deskriptive Ergebnisse der Algorithmen

Insgesamt zeigte Naive-Bayes bessere Ergebnisse als die Decisionlistsmethode und zwar sowohl mit ($t = 38,5$; $p<0,001$) als auch ohne Smoothing ($t = 4,5$; $p<0,001$).Weiterhin waren Naive-Bayes-Algorithmen mit Smoothing denjenigen ohne Smoothing überlegen ($t = -7,8$; $p<0,001$), während es sich bei Decisionlists umgekehrt verhielt ($t = 11,2$; $p<0,001$).

Aufgrund des unterschiedlichen Verhaltens der Methoden und der signifikanten Unterschiede bei Verwendung von Good-Turing wurde für jede Methode diejenige Gruppe mit dem besten Ergebnis ausgewählt und nachfolgende Analysen entsprechend für Naive-Bayes mit Smoothing und Decisionlists ohne Smoothing getrennt durchgeführt.

Bei varianzanalytischer Überprüfung der Bereiche Merkmal, Position und Distanz mittels des nicht-parametrischen Kruskall-Wallis-Verfahrens ließ sich für Naive-Bayes mit Smoothing ein Einfluß nur für Position ($\chi^2 = 105, 68$; $p < 0, 001$) und Distanz ($\chi^2 = 99,9$; $p<0,001$)

feststellen. Algorithmen, die die Position (*fein* und *grob*) miteinbezogen, waren hierbei jenen, die nur die Position (*bag*) betrachteten, deutlich überlegen ($U = 1578{,}5$; $p<0{,}001$).[3] Zudem war die Betrachtung nur des Diskursmarkersatzes (Distanz *klein*) allen anderen Distanzen überlegen.

Im Falle der Decisionlists zeigte sich nur die Art betrachteter Merkmale ($\chi^2 = 25{,}3$; $p<0{,}001$) als relevant. Am effektivsten erwiesen sich dabei `pos`-Tags, gefolgt von `tokens`. Die verschiedenen Distanzen zeigten keine signifikanten Unterschiede ($\chi^2 = 14{,}7$; $p = 0{,}25$).

5 Diskussion

Das Ziel der vorliegenden Pilot-Studie lag in der Überprüfung der Übetragbarkeit korpusbasierter Verfahren zur Polysem- und Homonym-Disambiguierung auf den Bereich der Diskursmarkerdisambiguierung. Dieser Versuch wurde anhand des Beispiels wenn durchgeführt. Des weiteren sollten erste Hinweise hinsichtlich der für diese Aufgabe relevanten Merkmale gewonnen werden.

Die Überlegenheit von Naive-Bayes läßt darauf schließen, daß es einer Kombination verschiedener Merkmale bedarf, um die Disambiguierung von Diskursmarkern erfolgreich vornehmen zu können. Übereinstimmend wiesen beide Methoden darauf hin, daß die Betrachtung der Position eines Merkmals im Kontext des Diskursmarkers eine wichtige Rolle bei der Disambiguierung von wenn spielt. Ähnlich deutliche Aussagen lassen sich für Distanz und Merkmale nicht machen, da sie sich in Abhängigkeit von der Methode als relevant oder nicht relevant herausstellten.

Keines der Verfahren erreichte, unabhängig vom verwendeten Merkmalsset, bessere Ergebnisse als durch Zuweisung der häufigsten Relation erreicht würde. Gründe hierfür könnten zum einen in der Vielfalt rhetorischer Relationen (Granularität) zu suchen sein, die durch den untersuchten Diskursmarker wenn markiert werden können. Wie Kilgarriff (1993) zeigen konnte, ist das Ausmaß an Granularität auch auf seiten der Annotation von Bedeutung, da sehr feine Unterscheidungen bereits für Menschen z.T. nur schwer nachvollziehbar sind, so daß mit Unstimmigkeiten bei der Annotation von Korpora per Hand zu rechnen ist. Dem entsprechend stellt eine sehr feine Differenzierung natürlich auch höhere Anforderungen an das Disambiguierungssystem. Die Unterschiede zwischen den markierten Relationen von wenn sind zum Teil recht subtil und könnten so, ähnlich wie im Bereich der Polysemdisambiguierung die Möglichkeiten statistischer Verfahren überschreiten (Kilgarriff 1992, 1993). Demgemäß sollte abgewogen werden, ob ein so hoher Granularitätsgrad wie in der vorliegenden Studie für eine bestimmte Aufgabe notwendig und sinnvoll ist. Zu klären bleibt in diesem Zusammenhang auch, wie 'feinteilig' die Unterscheidungen zwischen den Relationen sein dürfen bzw. sein müssen, um eine gute Differenzierung von Diskursmarkern zu ermöglichen. In einem nächsten Schritt wäre deshalb die Wahl von Diskursmarkern mit einer reduzierten Anzahl funktionaler Relationen anzuraten.

[3] Testung mit dem nicht-parametrischen Verfahren von Mann und Whitney (1947) (Mann-Whitney-U-Test).

Der Zugewinn an Präzision, der durch die Verwendung des Good-Turing-Verfahrens zum Ausgleich einer geringen Datengrundlage bei Naive-Bayes erzielt wurde, legt weiterhin nahe, daß bei Verwendung einer größeren Datenmenge bessere Ergebnisse zu erreichen wären. Darauf weisen auch Untersuchungen von Ng (1997) hin, die ergaben, daß die Güte von Disambiguierungsalgorithmen stark von der Anzahl verfügbarer Trainingsbeispiele abhängt. Demgemäß sollten in nachfolgenden Untersuchung wesentlich mehr Trainingsbeispiele eingesetzt werden, als dies im Rahmen dieser Pilot-Studie vorgesehen und möglich war.

Zusammenfassend läßt sich feststellen, daß die Übertragung daten-basierter Methoden der Polysem- und Homonymdisambiguierung auf den Bereich der Diskursmarker ohne größere Anpassungen möglich scheint. Wie allerdings die unbefriedigenden Ergebnisse im Rahmen dieser Pilot-Studie zeigen, müssen jedoch weitere Anstrengungen unternommen werden, um die Methoden effektiver einsetzbar zu machen. Zum einen ist hierfür die Erstellung größerer Korpora erforderlich, zum anderen könnte das Hinzuziehen externer Datenquellen wie Diskursmarkerlexika helfen, systematischere Informationen über die möglichen rhetorischen Spektren einzelner Diskursmarker zu erhalten. Die Erstellung solcher externen Ressourcen sollte demnach vorangetrieben werden. Wie weiterhin gezeigt wurde, reicht die Betrachtung eines einzelnen Merkmals zur Disambiguierung von wenn nicht aus. Vermutlich dürfte dies für die meisten komplexeren Diskursmarker in gleicher Weise gelten. Sinnvoll scheint es deshalb, noch genauer auf die jeweiligen Merkmalskombinationen einzugehen, die sich für die Unterscheidung der jeweiligen rhetorischen Relationen als relevant erweisen könnten. Die Testung weiterer Disambiguierungsverfahren (z. B. example-based Verfahren; Dagan/ Lee/Pereira 1997; Ng/Lee 1996) und der systematische Vergleich auf ihre Tauglichkeit für die vorliegende Aufgabe wäre ebenfalls zu begrüßen.

Danksagung

Mein Dank gilt Justin Salisbury und Csilla Puskas, die maßgeblich an der Aufbereitung des Korpusmaterials beteiligt waren, Georg Rehm für zahlreiche hilfreiche Hinweise und Anregungen zu Beginn der Arbeit und zwei anonymen Reviewern für wertvolle Anmerkungen zu einer früheren Version dieses Textes.

Literaturangaben

Agirre, Eneko/Martinez, David (2000): "Exploring automatic word sense disambiguation with decision lists and the web". In: *Proceedings of the COLING Workshop on Semantic Annotation and Intelligent Content*. Saarbrücken, Germany: 11-19.
See: http://arxiv.org/abs/cs.CL/0010024

Dagan, Ido/Lee, Lillian/Pereira, Fernando (1997): "Similarity-based methods for word-sense disambiguation". In: *Proceedings of the 35th ACL and the 8th EACL*. Madrid, Spain: 56-63.

Dipper, Stephanie/Kermes, Hannah/König-Baumer, Esther/Lezius, Wolfgang/Müller, Frank/ Ule, Tylman (2002): *Dereko. German reference corpus (Tech. Rep.)*. Universität Tübingen.

Domingos, Pedro/Pazzani, Michael (1996): "Beyond independence: Conditions for the optimality of the simple bayesian classifier". In: Saitta, Lorenza. (ed.): *Machine Learning: Proceedings of the Thirteenth International Conference*. San Francisco: 105-112.

Escudero, Gerard/Marquez, Llufs/Rigau, German (2000): "A comparison between supervised learning algorithms for word sense disambiguation". In: Cardie, Claire/Daelemans, Walter Nedellec,Claire/ Tjong Kim Sang, Erik (eds): *Proceedings of the 4th Conference on Computational Natural Language Learning, CoNLL 2000*. Lisbon, Portugal: 31-36.

Fellbaum, Christiane (1998): *WordNet: An eletronic lexical database*. Cambridge.

Good, I. J. (1953): "The population frequencies of species and the estimation of population parameters". *Biometrika* 40: 237-264.

Hovy, Eduard H. (1993): "Automated discourse generation using discourse structure relations". *Artificial Intelligence* 63: 341-385.

Ide, Nancy/Priest-Dorman, Greg (2000): *Corpus encoding standard (Tech. Rep.)*. Online available: www.cs.vassar.edu/CES/.

Ide, Nancy/Veronis, Jean (1998). "Word sense disambiguation: The state of the art". *Computational Linguistics* 24 (1): 1-41.

Kilgarriff, Adam (1992): *Polysemy*. Unpublished doctoral dissertation, University of Sussex.

Kilgarriff, Adam (1993): "Dictionary word sense distinctions: An enquiry into their nature". *Computers and the Humanities*: 26: 365-387.

Kurohashi, Sadao/Nagao, Makoto (1994). "Automatic detection of discourse structure by checking surface information in sentences". In: *Proceedings of the 15th International Conference on Computational Linguistics COLING 1994*. Kyoto, Japan: Vol. 2, 1123-1127.

Li, Xiaobi/Szpakowicz, Stan/Matwin, Stan (1995): "A WordNet-based algorithm for word sense disambiguation". In: Mellish, Chris S. (ed.): *Proceedings of the 14th International Joint Conference on Artificial Intelligence; Montréal, Québec, Canada*. San Mateo: 1368-1374.

Manning, Christopher D.,/Schütze, Hinrich (1999): *Foundations of statistical natural language processing*. Cambridge, Massachusetts.

Mann, Henry/Whitney, Donald (1947): "On a test of whether one of two varables is stochastically larger than the other". *Annals of Mathematical Statistics* 18: 50-60.

Mann, William C./Matthiessen, Christian M.I.M./Thompson, Sandra A. (1989): *Rhetorical structure theory and text analysis (ISI Research Report)*. University of Southern California.

Mann, William C./Thompson, Sandra A. (1988): "Rhetorical structure theory: Towards a functional theory of text organization". *Text* 8 (3): 243-281.

Marcu, Daniel (1998): "A surface-based approach to identifying discourse markers and elementary textual units in unrestricted texts". In: *COLING/ACL'98 Workshop on Discourse Relations and Discourse Markers*. Montreal, Canada: 1-7.

Marcu, Daniel (2000): *The theory and practice of discourse parsing and summarization*. Cambridge, London.

Miller, Georg A. (1990): "WordNet: An on-line lexical database". *International Journal of Lexicography* 3 (4): 235-312.

Millis, Keith K./Golding, Jonathan/Barker, Gregory (1995): "Causal connectives increase inference generation". *Discourse Processes* 20 (1): 29-50.

Mooney, Raymond J. (1996): "Comparative experiments on disambiguating word senses: An illustration of the role of bias in machine learning". In: Brill, Eric (ed.): *Proceedings of Empirical Methods in Natural Language Processing, University of Pennsylvania*. Philadelphia, Pa.: 82-91.

Müller, Frank (2002): *Shallow-parsing stylebook for german (Tech. Rep.)*. Universität Tübingen.

Ng, Hwee Tou (1997): "Getting serious about word sense disambiguation". In: *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semanics: Why, What and How?* Washington, USA: 1-7.

Ng, Hwee Tou/Lee, Hian Beng (1996): "Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach". In: *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics, Santa Cruz, CA*. Morristown, NJ: 40-47.

Stede, Manfred/Umbach, Carla (1998): "DiMLex: A lexicon of discourse markers for text generation and understanding". In: *Proceedings of COLING/ACL 1998*. Montreal, Canada: 1238-1242.

Ule, Tylman (2002): *Dereko linguistic markup (Tech. Rep.)*. Universität Tübingen.

Wilks, York/Stevenson, Mark (1997): "Combining independent knowledge sources for word sense disambiguation". In: Mitkov, Ruslan/Nicolov, Nicolas (eds.): *Proceedings of the Conference Recent Advances in Natural Language Processing,* Tzigov Chark, Bulgaria. Amsterdam: 1-7.

Wilks, York/Stevenson, Mark (1998): "Word sense disambiguation using optimised combinations of knowledge sources". In: *Proceedings of COLING-ACL98*. Montreal: 1398-1402.

Yarowsky, David (1992): "Word-sense disambiguation using statistical models of roget's categories trained on large corpora". In: Boitet, Ch. (ed.): *Proceedings of the 14th International Conference on Computational Linguistics, COLING'92*. Grenoble: 454-460.

Yarowsky, David (1994): "Decision lists for lexical ambiguity resolution: Application to accent restoration in spanish and french". In: *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, Las Cruces, New Mexico, USA*. Morristown, NJ: 88-95.

Yarowsky, David (1999): "Hierarchical decision lists for word sense disambiguation". *Computer and the Humanities* 34 (1 - 2): 179-186.

ISSN 1615-3014

Anhang A: RST-Relationen für *wenn*

Verwendete Relationen waren *background, concession, condition, contrast, dependent evaluation, elaboration, nonvolitional cause, progression, restriction, volitional cause, volitional result*.

Bis auf *dependent evaluation* und *restriction* entsprechen alle Relationen den in Mann et al. (1989) beschriebenen. Um Besonderheiten der Verwendung von wenn präziser fassen zu können, wurden die beiden Relationen *dependent evaluation* und *restriction* hinzugenommen. Eine Bedeutungsbeschreibung in Anlehnung an Mann et al. (1989) bietet Tabelle 3.

Relationsname:	DEPENDENT EVALUATION
Bedingungen an N:	Beinhaltet eine Bewertung
Bedingungen an S:	keine
Bedingungen an die Kombination von N+S:	Die in N präsentierte Bewertung gilt nur dann, wenn die Bedingungen in S wahr sind.
Effekt:	Der Leser erkennt, daß eine mögliche Situation oder Handlung in S durch N eine Bewertung erfährt.
Ort des Effekts:	N+S
Beispiel:	Es wäre ein Jammer, wenn er seinen Betrieb dicht machen müßte.

Relationsname:	RESTRICTION
Bedingungen an N:	die Situation oder Handlung wird negativ bewertet
Bedingungen an S:	die Situation oder Handlung in S soll negative Bewertung abschwächen
Bedingungen an die Kombination von N+S:	Die Situation oder Handlung in S erhöht den Wert der negativen Bewertung in N.
Effekt:	Die Bewertung der Situation oder der Handlung in N seitens des Leser verschlechtert sich.
Ort des Effekts:	N+S
Beispiel:	Wie aus internen Kreisen verlautet, sind nicht alle Antifagruppen von der neuen Strategie begeistert, auch wenn der demokratische Anstrich nur Mittel zum Zweck ist.

N: Nukleus; S: Satellit

Tabelle 3: Erläuterung zweier zusätzlicher rhetorischer Relationen