

Daniel Marcu & Abdessamad Echihabi  
(2002, University of Southern California)

# **An Unsupervised Approach to Recognizing Discourse Relations**

Martin Frey, Timo Taglieber

WS 07/08 · HS Diskursmodelle  
Seminar für Computerlinguistik  
Uni Heidelberg

## Motivation

### ▶ Beispiel:

A) *Such standards would preclude arms sales to states like Libya, which is also currently subject to a U.N. Embargo*

CONTRAST

B) *But states like Rwanda before its present crisis would still be able to legally buy arms.*

### ▶ Wie kann man Diskursrelationen maschinell erkennen?

▶ Wir suchen ein Verfahren zur (robusten) Klassifikation von Relationen zwischen willkürlich gewählten Textstellen

### ▶ Mögliche Ansätze:

▶ Diskursmarker/Cue Phrases → zu selten vorhanden

▶ Semantische Interpretation der Sätze (Wissenbasis) → nicht möglich

## An Unsupervised Approach to Recognizing Discourse Relations

# Ansatz

- ▶ Unüberwachtes Lernverfahren zum Klassifizieren von Diskursrelationen
- ▶ Beschränkung auf **vier Gruppen** von Relationen (vs. 25 Relationen in *Mann & Thompson, 1988*)
  1. CONTRAST (Diskursmarker „but“)
  2. CAUSE-EXPLANATION-EVIDENCE (Diskursmarker „because“)
  3. ELABORATION
  4. CONDITION
- ▶ Diskursmodelle können stark verschieden sein, haben aber diese vier Typen von Relationen gemeinsam
- ▶ Konsequenz: Klassifizierung wird gröber
- ▶ Vorteil: Einfaches Sammeln von „rauschfreien“ Daten dank Diskursmarker

## Ansatz – Gruppierung der Relationen

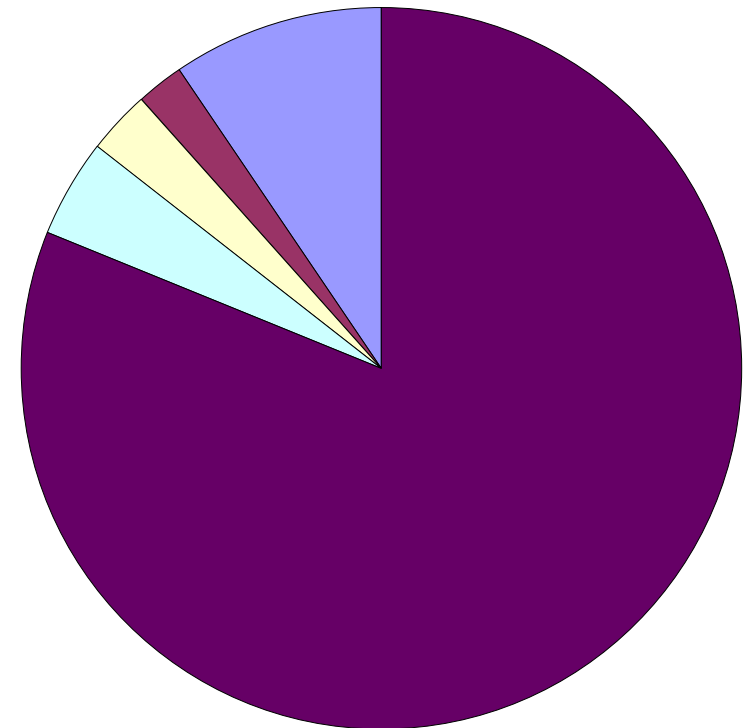
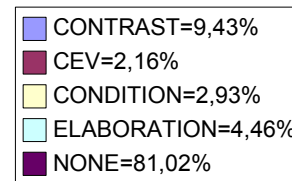
- ▶ CONTRAST nach Marcu und Echihabi umfasst folgende Klassen nach Mann und Thompson:
  - ▶ ANTITHESIS
  - ▶ CONTRAST
  - ▶ CONCESSION (Zugeständnis)
  - ▶ OTHERWISE

## Aufbau eines Trainingskorpus (1)

- ▶ Extraktion von gelabelten Paaren von Texteinheiten aus **Korpus „RAW“**
  - ▶ Entstand durch Verkettung mehrerer verfügbarer Korpora
  - ▶ Nicht annotiert
  - ▶ Rund 41 Mio. Sätze
- ▶ Unüberwachtes Mining mithilfe von Suchmustern für jede Relation  
z. B. Für die Relation CONTRAST (ungenau Methode!):
  - ▶ [`<satz>... </satz>`]      [`<satz> But ... </satz>`]
  - ▶ [`<satz> ... ]`                [`but ... </satz>`]
  - ▶ [`<satz> ... ]`                [`although ... </satz>`]
  - ▶ [`<satz> Although ... , ]`    [`... </satz>`]
  - ▶ Beide Teile in [ ] sind jeweils durch die Relation CONTRAST verbunden

## Ergebnisse der 1. Extraktion

- ▶ Nach dem Mining im RAW-Korpus (41.147.805 Sätze):



- ▶ Zusätzlich hinzugefügt:
  - ▶ 1.000.000 x NO-RELATION-SAME-TEXT
  - ▶ 1.000.000 x NO-RELATION-DIFFERENT-TEXTS
- ▶ Fertiger Trainingskorpus:
  - ▶ 9.811.574 Beispiele

# Kollokationen als Indikator für Diskursrelationen

- ▶ Hypothese: Bestimmte Wortpaare stehen häufig in bestimmter diskursiver Relation zueinander
  - ▶ John is **good** in math and sciences.  
Paul **fails** almost every class he takes.
  - ▶ „good“ und „fails“ deuten auf die Relation CONTRAST hin
  - ▶ Idee: Finden vieler solcher Wortpaare liefert robuste Methode zur Erkennung von Relationen
- ▶ Die Hypothese wird nun durch statistische Methoden getestet

## Ermittlung der Wortpaarwahrscheinlichkeiten

- Wir bilden zu jedem Beispiel-Paar  $(V, W)$  aus dem Trainingskorpus das kartesische Produkt über den Wörtern der beiden Texteinheiten:

$$\{v_i \in V\}$$

	$v_1$	$v_2$	...	$v_n$
$w_1$	$(v_1, w_1)$	$(v_2, w_1)$		
$w_2$		$(v_2, w_2)$		
...				
$w_n$				

$$\{w_j \in W\}$$

## Ermittlung der Wortpaarwahrscheinlichkeiten

- ▶ Relative Frequency Estimate für jedes Wortpaar über den Klassen:
  - ▶ Wie oft kommt das Wortpaar in jeder Relationsklasse vor
  - ▶ Bedingte Wahrscheinlichkeit für das Wortpaar  $(v_i, w_j)$  unter der Bedingung Relationsklasse  $r_k$ :
 
$$p((v_i, w_j) | r_k) = \frac{\text{Anzahl der Auftretenshäufigkeit } (v_i, w_j) \text{ in } r_k}{\text{Auftretenshäufigkeit des Wortpaares insgesamt}}$$
  - ▶ Zur Vermeidung von Nullwahrscheinlichkeiten: Smoothing für alle Paare, die irgendwo nicht vorkommen

## Welche Relation besteht zwischen den Texteinheiten?

- ▶ Berechnung der Wahrscheinlichkeit, dass zwei Texteinheiten  $V$  und  $W$  in einer bestimmten Relation  $r_k$  zueinander stehen
- ▶  $\operatorname{argmax}_{r_k} p(r_k \mid (V, W))$ 
  - ▶ Um zu erfahren, in welcher Relation zwei Texteinheiten zueinander stehen, berechnen wir ihre Wahrscheinlichkeit für jede einzelne Relation und nehmen dann davon die größte Wahrscheinlichkeit
- ▶ Mit dem Satz von Bayes können wir dafür einfach rechnen:
  - ▶  $\operatorname{argmax}(\log p((V, W) \mid r_k) + \log p(r_k))$
  - ▶  $p((V, W) \mid r_k) = \text{Produkt aller } v_i, w_j \in (V, W) \text{ über } p((v_i, w_j) \mid r_k)$

## Training der Wortpaar-basierten Klassifizierer (1)

- ▶ Auf diese Art und Weise werden für alle Paare von Relationen ( $r_a$ ,  $r_b$ ) Wortpaar-basierte Klassifizierer am Trainingskorpus trainiert
  - ▶ Die Cue-Phrases wurden dabei entfernt, sonst würde nur gelernt zwischen den Suchmustern der Relationen zu unterscheiden
- ▶ Jeder Klassifizierer wird an einem Korpus getestet:
  - ▶ 5.000 Beispiele von  $r_a$  / 5.000 Beispiele von  $r_b$
- ▶ Resultate sind vorzeigbar, hier die Top-3:
  - ▶ CEV/ELABORATION=93%
  - ▶ CONDITION/ELABORATION=89%
  - ▶ CONTRAST/CEV=87%
  - ▶ Untergrenze: 64% (Baseline 50%!)

## Auswertung der Resultate (1)

- ▶ Lernkurve flacht bei einigen Paaren ab:
  - ▶ Beim Klassifizierer CONTRAST/CEV steigt die Genauigkeit lediglich von 87,1% (bei 2,0 Mio. Beispielen) auf 87,3% (bei 4,7 Mio. Beispielen)  
Trotz des Verdoppelns der Trainingsbeispiele!
  
- ▶ Zur Erklärung folgende Hypothese:
  - ▶ Es gibt ein Grundrauschen im Trainingskorpus (Suchmuster teilw. versagt)
  - ▶ Zuviele der Wortpaare sind kaum relevant (siehe Kart. Produkt)
    - ▶ Als Indikatoren ungeeignet um Relationen vorherzusagen
  
- ▶ Test der Hypothese durch Wiederholen des Experiments

## Aufbau eines Trainingskorpus (2)

- ▶ Extraktion diesmal aus einem neuen **Korpus „BLIPP“**  
(Brown Laboratory for Linguistic Information Processing)
  - ▶ Annotiert (Parse-Bäume)
  - ▶ Rund 1,8 Mio. Sätze
- ▶ Beschränkung auf die repräsentativsten Wörter der Texteinheiten:
  - ▶ Nomina
  - ▶ Verben
  - ▶ Cue-Phrases
- ▶ Ergebnisse der 2. Extraktion vergleichbar mit RAW-Korpus
  - ▶ 435.690 Beispiele (inkl. 58.000 x 2 für NO-RELATION)

## Training und Auswertung der Resultate (2)

- ▶ Training der Wortpaar-basierten Klassifizierer analog zu RAW-Korpus
- ▶ Resultate sind deutlich schlechter, hier die Top-3:
  - ▶ CEV/ELABORATION=82%
  - ▶ CONTRAST/ELABORATION=78%
  - ▶ CONDITION/ELABORATION=78%
  - ▶ Untergrenze: 58% (Baseline 50%!)
- ▶ Lernkurve aber steiler:
  - ▶ Genauigkeit von 82% schon nach 100.000 Beispielen erreicht (mit RAW-Korpus: nach 1.000.000)

## Interpretation beider Experimente

- ▶ Die Resultate der Tests mit den Korpora RAW und BLIPP lassen schlussfolgern:
  - ▶ Die Beschränkung auf repräsentativste Wörtersowie
  - ▶ ein größerer (annotierter) Trainingskorpussollten den Bau von leistungsfähigeren Klassifizierern ermöglichen (um letztendlich Diskursrelationen maschinell besser erkennen zu können)

## An Unsupervised Approach to Recognizing Discourse Relations

## Quellen

- ▶ Introduction to RST, Mann & Taboada, 2007  
<http://www.sfu.ca/rst/01intro/intro.html>
- ▶ Marcu, D. and Echihabi, A. 2002. An unsupervised approach to recognizing discourse relations. In ACL-02, 368-375.  
<http://acl.ldc.upenn.edu/P/P02/P02-1047.pdf>

*Danke für eure Aufmerksamkeit!*