

An Unsupervised Method for Word Sense Tagging using Parallel Corpora

von Mona Diab und Philip Resnik

Souhail Bouricha
7. Februar 2011

Im Seminar „Word Sense Disambiguation“ bei Stefan Thater

Einleitung

- Fortschritte bei Wordsense Disambiguation durch:
 - Machine Learning
 - Wordnet
 - Große Corpora

Machine Learning

- Supervised Methods:
 - + gute Ergebnisse
 - teuer
 - langsam
- Unsupervised Methods:
 - relativ schlechte Ergebnisse
 - + billig

Beide Vorteile kombinieren?

- Parallele Corpora verwenden
 - (einen Text, der in zwei Sprachen vorliegt)
 - Vereint Vorteile:
 - bessere Ergebnisse als rein unsupervised methods
 - billiger als rein supervised methods
 - Können für WS Disambiguation verwendet werden, weil bei Übersetzungen Ambiguitäten entstehen

WSD mit parallelen Corpora

- Ein Wort, das in einer Sprache mehrere Bedeutungen hat, wird in der anderen Sprache im gleichen Kontext mit dem gleichen Wort übersetzt:

[FR] catastrophe → [EN] disaster, tragedy ...

Jedes dieser Worte (disaster ...) ist wieder ambig

Idee des Papers

- Unsupervised Algorithmus verwenden
- Zwei parallele Corpora
- Wordnet (= Sense-Inventar) gibt es nur für Englisch
- Wir finden für diesen Corpus die Bedeutungen mit Hilfe des Parallelcorpus in der anderen Sprache

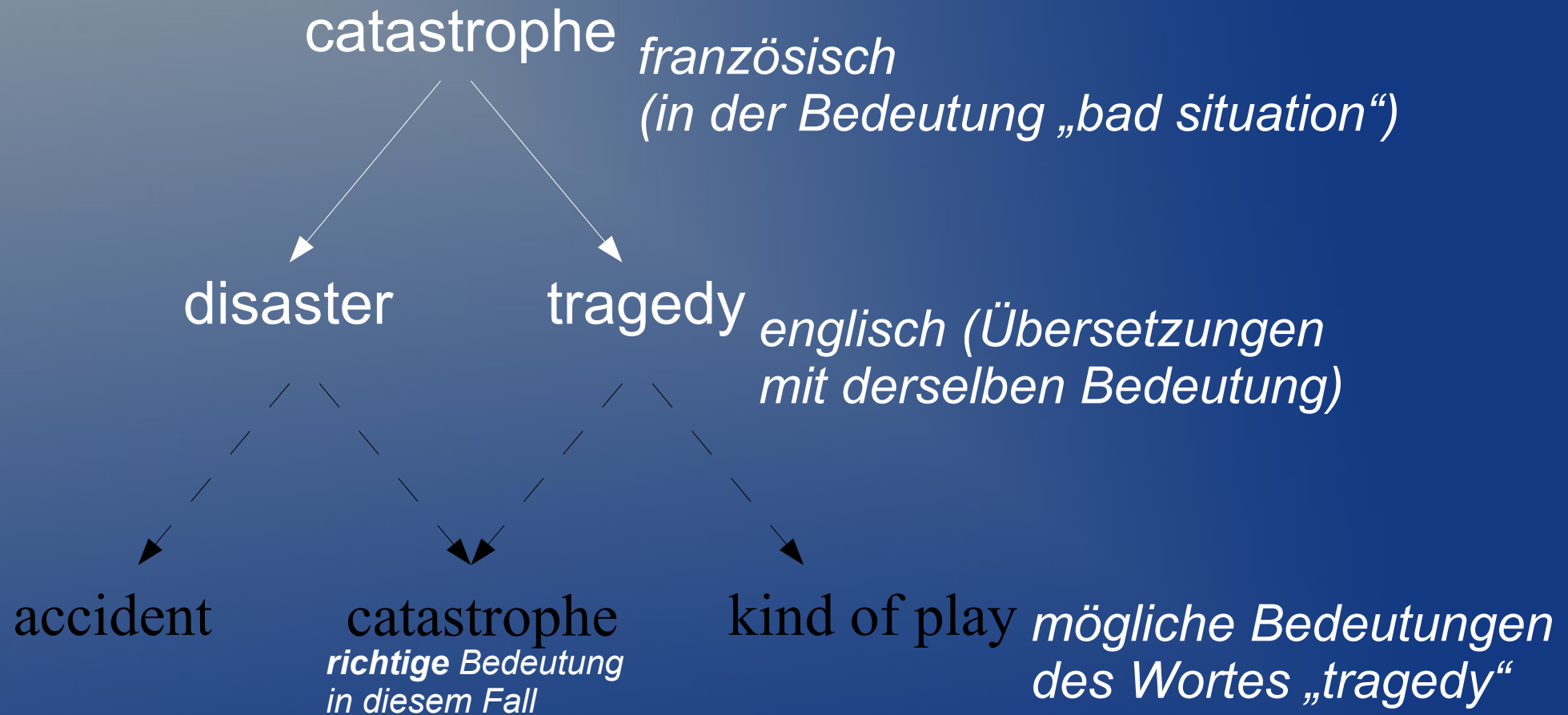
Zwei Hauptziele

- Große Menge von semantisch annotierten Daten produzieren, für die Sprache mit dem Sense-Inventar
- Wir verwenden das gleiche Sense-Inventar für die zweite Sprache
Das ermöglicht uns, einen sense-tagged Corpus zu erstellen

Zwei Faktoren

- Die Instanzen von einer Word/Sense-Kombination werden konsistent mit wenigen Wörtern in eine andere Sprache übersetzt
- Diese Wörter sind selten ein Singleton-Set

Beispiel (englisch – französisch)



Paper Approach

- Wir identifizieren Wörter in beiden Corpora und ordnen sie aneinander an
- Wir bilden ein Target-Set (Englisch) mit Wörtern, die die gleiche Übersetzung im französischen Text (Source) haben
- Wir suchen die gemeinsame Bedeutung von allen Wörtern im Target-Set
- Die gefundene Bedeutung projizieren wir dann zurück auf den französischen Text (Source)
- Wir evaluieren die Performanz an einem Gold-Standard-Text

Step 1: Alignment

Englischer Corpus (Target)

The chaos in the room ...

The disaster of the floods ...

The tragedy turned into ...

Französischer Corpus (Source)

La catastrophe dans la chambre ...

La catastrophe des inondations ...

La catastrophe est devenue ...

- Alignment ist keine triviale Aufgabe
- Jeder Übersetzer übersetzt anders.

Step 2: Mapping

Wir suchen für jedes französische Wort das englische Wort und bilden mit den gefundenen Wörtern eine Menge:

Englischer Corpus (Target)

The chaos in the room ...
The disaster of the floods ...
The tragedy turned into ...

Französischer Corpus (Source)

La catastrophe dans la chambre ...
La catastrophe des inondations ...
La catastrophe est devenue ...

Step 3: Sense Assignment

Wir suchen die gemeinsame Bedeutung von {chaos, disaster, tragedy}.

Jedes Wort hat mehrere Bedeutungen:

Chaos: 1. Accident 2. Bad situation 3. Catastrophe

Disaster: 1. Catastrophe 2. Accident

Tragedy: 1. Drama 2. Catastrophe

Wir wählen die Bedeutung, die alle Wörter (Chaos, Disaster, Tragedy) gemeinsam haben: Catastrophe

Englischer Corpus (Target)

The chaos [cat.] in the room ...

The disaster [cat.] of the floods ...

The tragedy [cat.] turned into ...

Step 4: Meaning projection

- Wir projizieren die Bedeutung, die wir für diese englischen Wörter (Synonyme) gefunden haben, auf den französischen Text (Source)

Englischer Corpus (Target)

The chaos [cat.] in the room ...
The disaster [cat.] of the floods ...
The tragedy [cat.] turned into ...



Französischer Corpus (Source)

La catastrophe [cat.] dans la chambre ...
La catastrophe [cat.] des inondations ...
La catastrophe [cat.] est devenue ...

Evaluation

- Wichtige Punkte in der Evaluation:
 - 1 Ein paralleler Corpus, mit Englisch auf einer Seite, groß genug, um ein stochastisches Übersetzungsmodell zu erzeugen
 - 2 Wir annotieren eine Untermenge des Corpus als Gold-Standard
 - 3 Wir brauchen Vergleichszahlen von anderen Systemen

Evaluation

- Problem: diese drei Bedingungen gleichzeitig zu erfüllen
 - Es gibt nur wenige und kleine menschlich-annotierte Corpora
 - Es gibt kein Gegenstück dazu (auch annotiert) in einer anderen Sprache

Evaluation

- Lösung des Problems:
 - Wir bilden einen großen englischen Corpus mit einem kleinen Teil annotierter Daten (Gold-Standard)
 - Übersetzen den ganzen Corpus mit MT-Technologie und erhalten einen künstlichen parallelen Corpus

Evaluation

- Künstlicher paralleler Corpus:
 - schlechtere Ergebnisse als menschliche Übersetzung
 - + Pseudo-Translations können viel leichter erzeugt werden

Data Aquisition

- Verwendete Corpora:
 - Brown Corpus, Wall Street Journal, Senseval-1 und 2 übersetzt
 - Übersetzung dieses Gesamt-Corpus in Französisch und Spanisch, mit zwei kommerziellen MT-Programmen (Globalink und Systran)

Ergebnis

- Performanz:
60 % Precision, 50 % Recall
 - Recall hat Probleme wegen Unambiguous Translations, wegen „monotoner“ Übersetzungen der Programme
 - Performanz schlecht wegen einfachem Algorithmus

Conclusion

- Sense-annotierte Daten in zwei Sprachen zu bekommen ist schwierig, aber man kann gute Resultate erwarten
- Es werden keine anderen WSD-Techniken am Target-Corpus verwendet

Fragen und Antworten

- Wie können wir bessere Resultate bekommen?
 - Durch Clustering und Eliminieren von Outliers
- Warum taggen wir nur Nomen?
 - In Wordnet haben die Nomen eine detaillierte Struktur
- Können einzelne Source-Words in Compound Nouns übersetzt werden?
 - Das wäre möglich, wenn es ein Wörterbuch für die Source-Language mit solchen Compound Nouns gäbe

Fragen und Antworten

- Warum gibt es ein Limit für die Satzlänge beim Anordnen (Alignment)?
 - Lange Sätze haben eine komplexere Struktur und die Alignment-Precision wird schlecht
- Welche Sprachen sollten wir benutzen?
 - Das hängt von der Verfügbarkeit der Daten ab, und davon wie ähnlich die Sprachen sind

Danke!

Noch mehr Fragen?