

Akquisition von Paraphrasen mit Hilfe Paralleler Korpora



Victor Gabriel Saiz Castillo
Department of Computational Linguistics
Heidelberg University, Germany

Was ist eine Paraphrase?

„Umschreibung eines sprachlichen Ausdrucks mit anderen Wörtern oder Ausdrücken“

Beispiel:

- Die deutsche Fußballnationalmannschaft
- Die National-Elf

Warum Paraphrasen-Erkennung?

Paraphrasen können nützlich sein für:

- Text-Zusammenfassung
- Spracherkennung und Sprachsynthese
- Information Retrieval
- Statistische maschinelle Übersetzung
- Frage-Antwort-Systeme
- Suchmaschinen Optimierung

“To decide whether two sentences are paraphrases may well be the grand challenge for natural language processing.”

(Dr. Bill Dolan, Microsoft Research)

Was wird gebraucht um Paraphrasen zu Gewinnen?

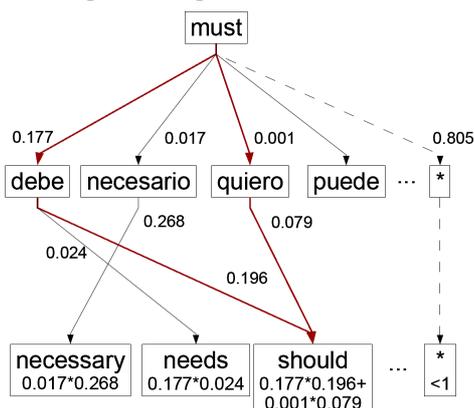
Um Paraphrasen zu Akquirieren werden in dieser Arbeit folgende Zutaten gebraucht:

1. Multilinguales alignierten Korpus (Europarl)
2. Programm für Alignment Erstellung und Berechnung der daraus resultierenden Übersetzungswahrscheinlichkeiten (GIZA++, Moses)

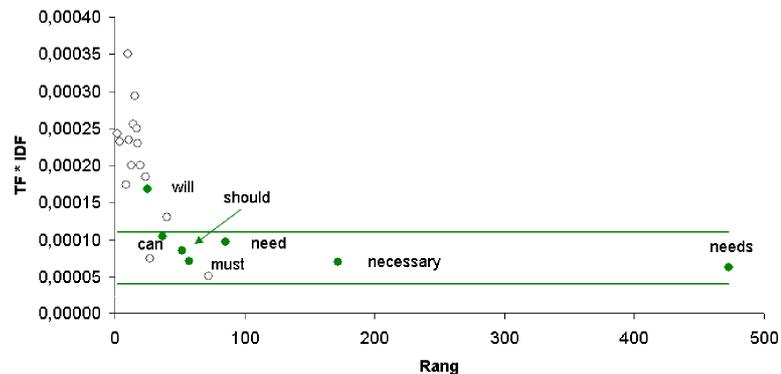
Wie kann man Paraphrasen Gewinnen?

Drei Verschiedene Klassifikatoren im Zusammenspiel:

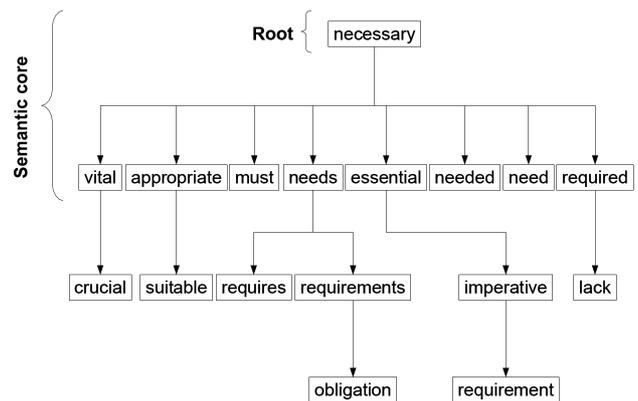
1) Bewertung aus Alignment-Wahrscheinlichkeiten



2) Relevanz-Korrelation



3) Überlappung mit einem semantischen Kern



Wie gut funktioniert das Verfahren?

Ergebnis-Vergleich für: “UNDER CONTROL”

Callison-Burch und Bannard:

checked, curb, curbed, in check, limit, slow down.

Saiz Castillo:

to monitor, under the control, controlled from, under the full scrutiny, to be controlled from, well under control, within control, be controlled from, dominated, monitor, in check, inspected, verified.

Autor	Umgebung	Alignments	Genauigkeit
Ibrahim et. al.	Monolingual	Automatisch	41,2
Callison-Burch und Bannard	Multilingual	Automatisch	65,4
Callison-Burch und Bannard	Multilingual	Manuell	84,7
Barzilay und McKeown	Monolingual	Automatisch	86,5
Saiz Castillo	Multilingual	Automatisch	96,2