

Satzalignierung

Alles über XAlign

28. November 2007

**Korpora und Korpusbearbeitung für die
maschinelle Übersetzung**

Lukas Bulwahn

bulwahn@in.tum.de

Überblick

- **theoretischer Teil:**
 - **Allgemeines über XAlign**
 - **verwendeten Algorithmen und Funktionsweise**
- **praktischer Teil**

Was ist XAlign?

- ein Satzalignierungswerkzeug
- für viele Sprachen verwendbar
- im Rahmen vom Projekt "Langue et Dialogue" in Nancy entstanden
- entwickelt von Patrice Bonhomme, Thi Minh Huyen Nguyen, und Sean O'Rourke
- geschrieben in Java, frei zugänglich und Teil von Unitex 2.0

Ziel bei XAlign

- **Bestehende Ansätze zur Satzalignierung nutzen, aber möglichst sprachunabhängig gestalten**
- **Keine gemeinsame Auffassung von einem "Wort"**
- **Keinerlei Wörterbuch oder morphologisches System notwendig**

Verwendete Algorithmen

- **structure-driven document alignment**
- **zwei Kategorien von Alignierungsalgorithmen:**
 - structural alignment: Korelation der Länge der Sätze ausnutzen
z.B. Brown, Church&Gale
 - lexical alignment: Korelation der Frequenz von Wortpaaren ausnutzen
z.B. Kay&Röscheisen, Fung&Church, Fung&McKeown

Funktionsweise von XAlign

- **Zuerst hierarchische Struktur des Textes ausnutzen**
- **Dann Feedback-Schleife von structural und lexical alignment:**
 - die lexikalische Analyse beeinflusst Kostenfunktion der strukturellen Analyse
 - die strukturelle Analyse beeinflusst die Distanzfunktion des Textes

Church&Gale Algorithmus

- **Problem: Church&Gale Algorithmus nimmt an, dass die Länge alignierter Sätze gleich ist**
- **sicherlich korrekt für europäische Paare, jedoch nicht für Chinesisch – Englisch**
- **Lösung: Normierung der Satzlängen für die beiden Sprachen**

Praktischer Teil

- **Installation von Unitex 2.0**
 - Dateien auf Michaela Geierhos Webseite zu finden
 - benötigt Java 1.6 (und g++ unter Linux)
- **Text kann entweder unstrukturiert oder im TEI Format vorliegen**

Demo