Master-Arbeit von Andrea Fritz (Herbstsemester 2016)

Titel:
Erstellung eines parallelen Arzneimittelinformations-Korpus (Deutsch-Französisch) und Optimierung von dafür einsetzbaren Part-of-Speech-Taggern

Abstract
The goal of this master thesis was to build a parallel corpus for information on medical use of drugs in German and French and to design and optimize PoS-Taggers that can be used to annotate such a corpus. We discuss the steps necessary to build the parallel corpus and describe the decisions we made when designing the taggers. In addition, we provide and interpret the most important numbers of the established parallel corpus. We demonstrate that we can process an already existing XML-File with drug information in order to get a parallel corpus with sentence alignments. Despite a limited set of domain-specic training data, we demonstrate that it is possible to successfully design a PoS-Tagger with wich we can annotate our parallel corpus and thus make a parallel corpus from the medical domain available.

Zusammenfassung
In dieser Masterarbeit widmen wir uns dem Aufbau eines parallelen Arzneimittel-korpus (Deutsch-Französisch) und der Konzipierung bzw. Optimierung von dafür einsetzbaren PoS-Taggern. Wir erörtern die Schritte, die für den Aufbau des parallelen Korpus nötig sind und schildern die Entscheidungen, die wir bei der Konzipierung der Tagger treffen. Darüber hinaus liefern und interpretieren wir die wichtigsten Kennzahlen des aufgebauten parallelen Korpus. Wir zeigen, dass wir eine bestehende XML-Datei mit Arzneimittelinformationen so aufbereiten können, dass wir ein paralleles Korpus mit Satzalignierung erhalten. Weiter demonstrieren wir, dass wir trotz beschränktem domänenspezischem Trainingsmaterial Tagger konzipieren können, anhand derer wir das Arzneimittel-Korpus erfolgreich mit PoS-Tags versehen und so ein linguistisch annotiertes paralleles Korpus aus dem medizinischen Bereich bereitstellen können.