Erschienen in: Blühdorn, Hardarik/Breindl, Eva/Waßner, Ulrich W. (Hrsg.): Text - Verstehen. Grammatik und darüber hinaus. – Berlin, New York: de Gruyter, 2006. S. 411-415. (Institut für Deutsche Sprache. Jahrbuch 2005), https://doi.org/10.1515/9783110199963.bm

Stefan Klatt

# Kombinierbare Textanalyseverfahren für die Korpusannotation und Informationsextraktion

## 1. Einleitung

Im Rahmen dieser Arbeit wurden Präzisionsverfahren zur Textanalyse (im Besonderen für die Korpusannotation und die Informationsextraktion) für das Deutsche entwickelt, die auch autonom eingesetzt werden können.

Vom Anfang der Textanalyse bis hin zur Syntaxanalyse werden bestehende Probleme relevanter Verarbeitungsschritte diskutiert und Lösungswege zu deren Behebung dargeboten. Neben Problemen innerhalb der Verarbeitungsschritte werden dabei auch sonst gern vernachlässigte Probleme im Schnittstellenbereich zwischen den Verarbeitungsschritten sowie bei der Ergebnisausgabe behandelt.

Bei den entwickelten Werkzeugen handelt es sich um einen Tokenizer, einen regelbasierten Part-of-Speech-Tagger und einen mehrstufigen Parser, die alle mittels der Analysetechnik Pattern-Matching Easy-First Planning (PEP) implementiert wurden. Weiterhin wurden zwei einfach aufgebaute, rein korpusbasierte Verfahren zur Interpretation unbekannter Wörter und zur Extraktion fremdsprachlichen Materials entwickelt.

Eine Evaluierung der mit PEP entwickelten Werkzeuge (Tokenizer, Tagger und partieller Parser) führte zu besseren Resultaten gegenüber allen anderen bekannten Systemen in diesen Bereichen. Aber auch die beiden korpusbasierten Verfahren liefern ansprechende Ergebnisse und interessante Einsichten.

## 2. Die Analysetechnik PEP

Im Folgenden werden kurz die beiden wichtigsten Konzepte der Analysetechnik PEP vorgestellt. Alle weiteren Merkmale können Klatt (2005) entnommen werden.

### 2.1 Pattern-Matching als Analysekonzept

Durch Anwendung von Suchmustern als Bestandteil deterministischer Automaten (vorrangig die Muster *nb* und *seq*) auf eine gegebene natürlichsprachliche Äußerung (in der Regel ein Satz) wird für diese Äußerung eine vorgegebene Aufgabe durchgeführt.

Das n-stellige Suchmuster *nb* untersucht die Äußerung auf *n* benachbarte Elemente, die vorgegebene Bedingungen erfüllen. Falls eine solche Konstellation existiert (in (1) die Abfolge Determinativ – attributives Adjektiv – Nomen), kann das daraus resultierende Ergebniskonstrukt den bislang erkannten Konstrukten hinzugefügt werden und/oder können Teile davon nach vorgegebenen Operationen bearbeitet werden.

Das 2-stellige Suchmuster *seq* ist durch eine linke und eine rechte Grenze (*LB* und *RB*) definiert. Optional können dem Zwischenraum von *LB* und *RB* noch zu erfüllende Bedingungen auferlegt werden. Falls das angegebene Sequenzmuster existiert (in (2) die Sequenz mit LB = finites Auxiliarverb und RB = Vollverb im Partizip Perfekt), kann mit diesem wie zuvor unter *nb* beschrieben verfahren werden.

(1) Er besuchte [*NB1=DET* das] [*NB2=ADJA* Heidelberger] [*NB3=NN* Schloss].
(2) Er [*LB=VAFIN* hat] sein Herz in Heidelberg [*RB=VVPP* verloren].

### 2.2 Easy-First als Suchstrategie

Mit dem Begriff *Easy-First* sollen hier zwei Analyseprinzipien zum Ausdruck kommen. Zum einen werden Analyseentscheidungen in der Reihenfolge ihrer Einfachheit getroffen. Zum anderen besteht die Möglichkeit, sich mit einer

gefundenen Teilanalyse bzw. Grobanalyse (zunächst) zufriedenzugeben und die Suche nach Alternativlesarten in diesem Bereich aufzugeben bzw. zurückzustellen, so dass der zu betrachtende Suchraum und damit auch die Anzahl ambiger Teilanalysen teils drastisch reduziert werden kann.

## 3. Die Textanalyseverfahren im Überblick

### 3.1 Mit PEP implementierte Verfahren

Der PEP-Tokenizer besitzt eine wesentlich größere Funktionalität als herkömmliche Tokenizer. Letztere zerlegen einen Text in der Regel nur in Sätze und Wörter. Der PEP-Tokenizer hingegen markiert beispielsweise auch Äußerungseinheiten (bestehend aus einem oder mehreren Sätzen), Matrixsätze und Satzparenthesen, sofern sich diese mit Hilfe spezifischer Satzzeichenkonstellationen erkennen lassen. Weiterhin werden bestimmte Rechtschreibfehler (z.B. vergessene Leerzeichen vor Interpunktionszeichen) korrigiert sowie Problemfälle und unsichere Entscheidungen markiert. Letzteres ermöglicht eine schnelle manuelle Überprüfung mit dem Ziel, nahezu fehlerfrei tokenisierte Texte zu erstellen.

Diese Zielsetzung wurde eindrucksvoll durch ein Experiment bestätigt, bei dem die ersten 3000 mit einem Punkt endenden Tokens (Abkürzungen, Ordinalzahlen oder Satzendezeichen) eines bearbeiteten Korpus evaluiert wurden. Der dabei erzielte F(beta=1)-Score lag bei 99.87%. Da die verbliebenen fehlerhaften Entscheidungen allesamt als problematische Fälle erkannt und entsprechend markiert wurden, betrug der F(beta=1)-Score nach der manuellen Korrektur dieser Textstellen sogar 100%.

Der PEP-Tagger ist ein regelbasierter Part-of-Speech-Tagger, der in geringem Maße auch korpusbasierte Anfragen in den Disambiguierungsprozess integriert. Er benötigt keinerlei Trainingsdaten und besteht aus 18 Stufen, in denen sukzessive falsche kategoriale Lesarten von Wörtern ausgefiltert werden. Dabei werden in den ersten Stufen ungrammatische und sehr unwahrscheinliche Konstellationen betrachtet. In den nachfolgenden Stufen werden dann auch diverse unsichere Disambiguierungsentscheidungen getroffen.

Bezüglich des REFD-Korpus konnte nach Anwendung aller Stufen ein Recall von 99.65% und eine Precision von 97.50% erzielt werden. Die Disambiguierung der Restambiguitäten mittels des statistisch basierten TreeTaggers (Schmid 1995) führte zu einem F(beta=1)-Score von 99.51%, was einer Fehlerreduktion von über 50% gegenüber der alleinigen Anwendung des TreeTaggers entspricht.

Der PEP-Parser besteht aus drei Stufen. In der ersten Stufe wird ein Satz in dessen topologische Felder (siehe Höhle 1986) zerlegt. In der zweiten Stufe erfolgt in jedem topologischen Feld eine partielle syntaktische Analyse in sogenannte minimale Phrasen. In der dritten Stufe werden die minimalen Phrasen innerhalb eines Feldes in eine Konstituentenstruktur überführt, be-

vor dann die topologischen Felder selbst in eine Satzkonstituentenstruktur überführt werden.

Wiederum für das REFD-Korpus konnte für über 95% der Sätze eine korrekte Zerlegung in deren topologischen Felder vorgenommen werden. Für die ersten 500 Sätze des REFD-Korpus wurden für die Erkennung minimaler Phrasen folgende F(beta=1)-Score-Werte erzielt: Minimale Nominalphrasen 97.42%, minimale Determinativphrasen 98.69% und minimale Präpositionalphrasen 98.11%.

### 3.2 Restliche Verfahren

Des Weiteren wurden noch zwei korpusbasierte Verfahren zur Extraktion von Wortarteninformationen entwickelt, die hierfür nur ein tokenisiertes Korpus benötigen.

Das erste Verfahren zur Interpretation unbekannter Wörter basiert auf Häufigkeitsverteilungen signifikanter Kontexte, die auf das Vorliegen bestimmter Lesarten schließen lassen.

Das Verfahren wurde an drei unterschiedlich zusammengestellten Testmengen erprobt. Eine diesbezüglich vorgenommene Evaluierung bezüglich der Maße Precision und Recall lieferte nahezu gleiche Ergebnisse. Folglich ist zu erwarten, dass für die Interpretation weiterer unbekannter Wortformen ähnliche Ergebnisse erzielt werden. Die Erkennung unbekannter Adjektive und Verben weist dabei eine Korrektheitsrate von nahezu 100% auf, die erzielten Recall- und Precision-Werte von über 80% bei gewöhnlichen Nomina und Eigennamen sind ebenfalls mehr als zufriedenstellend.

Das zweite vorgestellte Verfahren extrahiert fremdsprachliche Wörter aus einem Korpus durch aktive Suche mittels eines einfachen Bootstrapping-Ansatzes. Dadurch ist es möglich, viele fremdsprachliche Wörter zu extrahieren, die entweder nur eine aus der Fremdsprache stammende Lesart besitzen oder darüber hinaus auch noch einheimische Lesarten aufweisen.

## 4. Anwendungsmöglichkeiten

Die in dieser Arbeit vorgestellten Textanalyseverfahren können sowohl autonom verwendet als auch miteinander kombiniert werden. Letzteres ermöglicht den Bau von Analysesystemen für diverse Anwendungen: von Systemen für die klassische Informationsextraktion (Jiao 2005) bis hin zu computerlexikographischen Systemen für die Extraktion von Verb-Nomen-Kollokationen (u. a. mit Hilfe topologischer Felderinformationen), Subkategorisierungsrahmen von Verben (mittels der ersten beiden Parsingstufen) oder Mehrwortausdrücken (bevorzugt in Vorfeldern). Mittlerweile wurde mit PEP auch ein Named-Entity-Recognizer für deutsche Texte implementiert (Ritz 2004).

Des Weiteren können mit den Verfahren hochqualitative Textkorpora erstellt werden mit dem Ziel, eine wertvolle Trainingsressource für entsprechende statistische Verfahren zu sein. Hierzu lassen sich die vollständigen bzw.

partiellen Analyseergebnisse des PEP-Parsers auch für die Erstellung syntaktisch annotierter Korpora nutzen (siehe Radeschütz 2005).

Damit existiert eine Reihe von Anwendungsmöglichkeiten, für die die Analysetechnik PEP und die zuvor vorgestellten Analyseverfahren wertvolle Dienste leisten können.

## Literatur

Höhle, Tilman N. (1986): Der Begriff ‚Mittelfeld'. Anmerkungen über die Theorie der topologischen Felder. In Weiss, Walter/Wiegand, Herbert Ernst/Reis, Marga (Hg.): Textlinguistik contra Stilistik? Wortschatz und Wörterbuch. Grammatische oder pragmatische Organisation von Rede? (= Kontroversen, alte und neue 3). Tübingen: Niemeyer. S. 329–340.

Jiao, Jie (2005): Specification and Implementation of an Information Extraction System for Crime Reports. Studienarbeit, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Klatt, Stefan (2005): Kombinierbare Textanalyseverfahren für die Korpusannotation und Informationsextraktion. Aachen: Shaker. (Dissertation, Universität Stuttgart).

Radeschütz, Sylvia (2005): Entwurf und Implementierung eines Annotationswerkzeugs für die Erstellung von Baumbanken. Diplomarbeit, Institut für Intelligente Systeme, Universität Stuttgart.

Ritz, Julia (2004): Entwurf und Implementierung eines Verfahrens zur Erkennung von Named Entities in einem Informationsextraktionssystem für die deutsche Sprache. Studienarbeit, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Schmid, Helmut (1995): Improvements in Part-of-Speech Tagging with an Application to German. In: Tzoukermann, Evelyne/Armstrong, Susan (Hg.): From Texts to Tags: Issues in Multilingual Language Analysis. Proceedings of the ACL Sigdat Workshop, Dublin, Ireland, ACL. S. 47–50.