TTLab Preprocessor – Eine generische Web-Anwendung für die Vorverarbeitung von Texten und deren Evaluation

Rüdiger Gleim und Alexander Mehler

Goethe-Universität Frankfurt

1 Einführung und Motivation

Dieser Beitrag stellt den *TTLab Preprocessor* (kurz: *TTLab PrePro*) als generische Web-Anwendung für die Vorverarbeitung von Texten in den *Digital Humanities* vor. Er erörtert die Architektur des *TTLab PrePro*, exemplifiziert das von ihm anvisierte Nutzungsszenario und fasst seinen aktuellen Entwicklungsstand zusammen.

Die linguistische Vorverarbeitung von Texten ist ein integraler Bestandteil jeder automatischen Textanalyse. Dies beinhaltet unter anderem die Erkennung der dem jeweiligen Text zugrundeliegenden Sprache(n), die Erkennung seiner logischen Dokumentstruktur, die Tokenisierung und Lemmatisierung seiner lexikalischen Konstituenten und die Annotation ihrer Wortarten (*PoS-Tagging*). Es existiert eine Reihe von Software-Systemen und -Komponenten, welche die Vorverarbeitung für verschiedene Sprachen umsetzen. In der Literatur werden dabei etwa für das PoS-Tagging Erkennungsraten von über 95% dokumentiert.[1] Für viele Fragestellungen, wie z.B. die Textklassifikation, fällt eine entsprechende Fehlerquote von ca. 5% kaum ins Gewicht. Im Bereich der *Digital Humanities*, bei der es etwa um die qualitative Analyse einzelner Wortbedeutungen geht, sind jedoch bereits Fehlerquoten von 1% oftmals inakzeptabel.[2] Gerade in diesem Bereich ist die automatische Vorverarbeitung zumeist der Ausgangspunkt für die nachfolgende unabdingbare manuelle Korrektur der Annotationen.

So stellt sich die Frage etwa zu Beginn eines Forschungsprojekts, wie hoch die erwartete Fehlerquote für Texte der untersuchten Sprache beim Einsatz eines bestimmten Präprozessierers ist. Zur Beantwortung dieser Frage kann eine Sammlung von Texten manuell vorverarbeitet und als so genannter *Gold-Standard* zur Bewertung der automatischen Vorverarbeitung herangezogen werden. Vergleicht man die Annotationsergebnisse verschiedener Systeme mit einem solchen Goldstandard, so können Kennzahlen zur Ermittlung der erwarteten Fehlerrate gewonnen werden, um schließlich den Aufwand für entsprechende manuelle Korrekturen zu schätzen. Da die Parametrisierung sowie die Ein- und Ausgabeformate verschiedener Systeme zur Vorverarbeitung variieren, ist die Durchführung einer solchen Evaluation aufwendig und ihrerseits fehleranfällig. Die Funktion, verschiedene Systeme über eine generische Schnittstelle nicht nur verwendbar, sondern auch evaluierbar zu machen, bildet folglich den funktionalen Kern des *TTLab PrePro*.

[1] Diese Rate schwankt erwartungsgemäß je nach Sprache und Genre der untersuchten Texte [Giesbrecht and Evert, 2009].

[2] Anne Bohnenkamp-Renken (2013); *persönliche Kommunikation*.

2 TTLab Preprocessor Web-Anwendung

Der *TTLab PrePro* ermöglicht die Vorverarbeitung von Texten, die automatische Evaluation auf der Basis von Goldstandards und die einzelfallbezogene Fehleranalyse. Die Eingabe in das System kann direkt über den Browser in Form einer Texteingabe, die Angabe einer Webressource oder den Upload von Dateien erfolgen. Die Upload-Funktion ermöglicht nicht nur das Hochladen mehrerer Dateien auf einmal, sondern auch die Verwendung von komprimierten Archiven. An Dateiformaten werden unter anderem HTML, PDF, RTF und DOC unterstützt. In der Voreinstellung wird die Sprache der Inputtexte automatisch erkannt und der für die jeweilige Zielsprache voreingestellte Präprozessierer verwendet. Es ist auch möglich, diese Parameter explizit zu setzen. Die Ausgabe erfolgt mittels *TEI P5* [TEI, 2014]. Die Ergebnisse können direkt im Browser in verschiedenen Sichten betrachtet und frei heruntergeladen werden.

Werden TEI-P5-Dokumente als Eingabe verwendet, so werden diese vom System – wie bei jedem anderen Eingabeformat – auf den unstrukturierten Text heruntergebrochen. Anschließend werden sie durch den Präprozessierer vorverarbeitet und in TEI P5 repräsentiert. Bilden annotierte TEI-P5-Dokumente den Input, so können diese als Goldstandard interpretiert werden. Das System evaluiert in diesem Falle den jeweils ausgewählten Präprozessierer auf der Basis dieses Goldstandards. Da die Tokenisierung zwischen den zu vergleichenden Dokumenten variieren kann, wird zunächst mittels dynamischer Programmierung ein Alignment der Token durchgeführt. Anschließend wird das Ergebnis der Lemmatisierung sowie des Taggings mit dem Goldstandard verglichen. Auf diese Weise können die aus dem *Machine Learning* bekannten Maße *Precision*, *Recall* und *F-Score* berechnet werden. Die Ergebnisse werden direkt im Browser angezeigt – sowohl für die einzelnen Dokumente, als auch für das Eingabekorpus insgesamt. Analog wird eine Rangverteilung der häufigsten Tagging- und Lemmatisierungsfehler (nach abnehmender Häufigkeit) visualisiert. Schließlich können die Tagging- und Lemmatisierungsfehler in einer tabellarischen Ansicht im jeweiligen Satzkontext untersucht werden. Abbildung 1 exemplifiziert eine solche Ansicht von Evaluationsergebnissen. Die obere Tabelle beinhaltet eine Liste aller evaluierten Dokumente mit den Gesamtergebnissen. Für ein ausgewähltes Dokument können, wie in diesem Beispiel gezeigt, Belegstellen von Tagging-Fehlern im Satzkontext aufzeigt werden. Dies erlaubt das gezielte Nachverfolgen und Beheben von Fehlern.

Der *TTLab PrePro* ist als Java- und JavaScript-basierte Client-Server-Architektur implementiert. Die Benutzeroberfläche ist mithilfe des JavaScript-Frameworks ExtJS realisiert. Das in *Apache Tomcat* laufende *Java Servlet* bearbeitet die Nutzeranfragen, ruft externe Systeme zur Vorverarbeitung auf, führt ggf. Evaluationen durch und bereitet die Ergebnisse für die Darstellung im Browser auf. In der aktuellen Version sind zwei Systeme des *TTLab Preprocessor* [Mehler et al., 2015, Waltinger, 2010] integriert sowie das System namens *Stanford CoreNLP* [Manning et al., 2014].

3 Zusammenfassung und Ausblick

Der vorliegende Beitrag stellt den *TTLab PrePro* als System zur Vorverarbeitung von Texten und darauf basierenden Evaluationen vor. Das mit der geplanten Publikation veröffentlichte System ist frei ver-

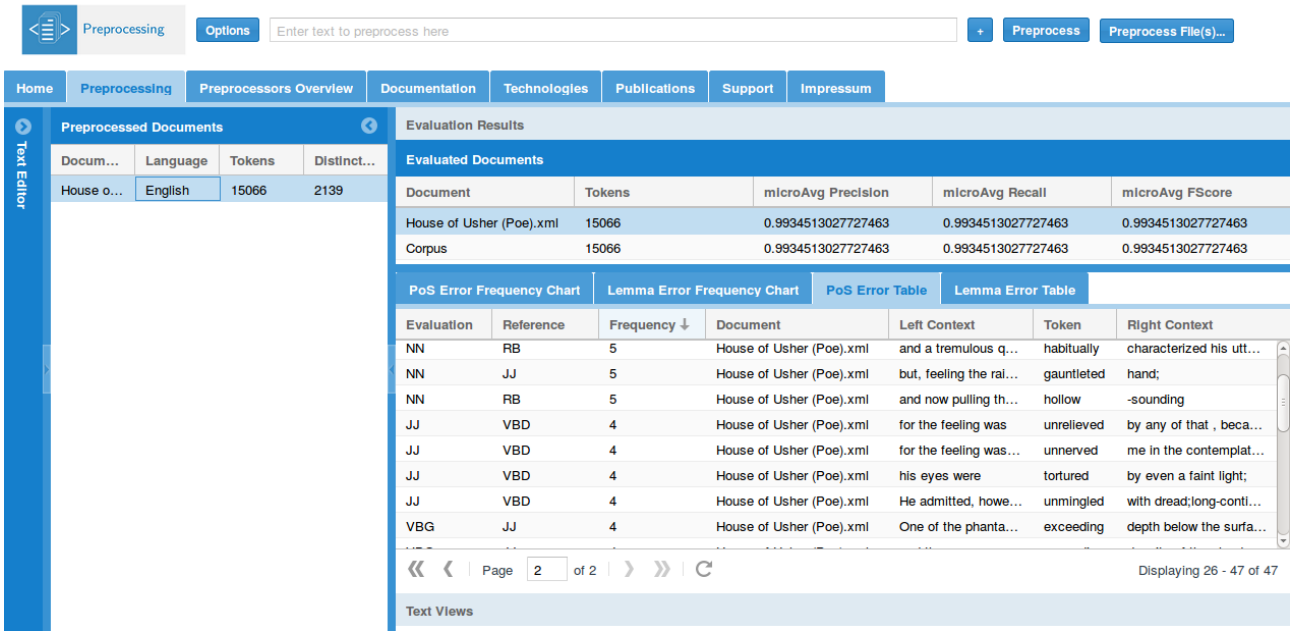

Abbildung 1: Ansicht von Evaluationsergebnissen, welche Tagging-Fehler mit Belegstellen im Satzkontext aufzeigt.

wendbar (*open access*). Die Weiterentwicklung zielt auf die Nutzbarmachung des UIMA-Frameworks[3]. Zum einen, um den Pool der verfügbaren Systeme zur Vorverarbeitung zu vergrößern, zum anderem, um umfangreiche Parameterstudien über die einzelnen Komponenten durchführen zu können. Ferner soll eine Normalisierung von PoS-Tagsets für die Evaluation entwickelt werden. Der *TTLab PrePro* zielt vor allem darauf, von Geisteswissenschaftlerinnen und -wissenschaftlern auch ohne Informatik-Vorkenntnisse genutzt werden zu können. Unterstützt werden derzeit die Sprachen Latein [Mehler et al., 2015], Englisch und Deutsch.

Der *TTLab PrePro* kann unter der URL `http://prepro.hucompute.org` getestet werden. Ein TEI-P5-Dokument zum Testen der Evaluation steht unter der Adresse `http://prepro.hucompute.org/examples/poe.tei` bereit.

Danksagung

Diese Arbeit ist im Rahmen des BMBF-Projekts *Computational Historical Semantics* (`www.comphistsem.org`) entstanden, für dessen Unterstützung wir uns herzlich bedanken.

Literatur

TEI P5: Guidelines for electronic text encoding and interchange, 2014. URL `\url{http://www.tei-c.org/Guidelines/P5/}`.

[3] `https://uima.apache.org/`

Eugenie Giesbrecht and Stefan Evert. An evaluation of part-of-speech taggers for the web as corpus. In *Proceedings of DGfS-CL Postersession 2009*, 2009.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, 2014. URL `http://www.aclweb.org/anthology/P/P14/P14-5010`.

Alexander Mehler, Tim vor der Brück, Rüdiger Gleim, and Tim Geelhaar. *Towards a Network Model of the Coreness of Texts: An Experiment in Classifying Latin Texts using the TTLab Latin Tagger*. Theory and Applications of Natural Language Processing. Springer, Berlin/New York, 2015.

Ulli Waltinger. *On Social Semantics in Information Retrieval*. Phd thesis, Bielfeld University, Germany, 2010.