GEFÖRDERT VOM

Universität Tübingen
Seminar für Sprachwissenschaft (SfS) Lehrstuhl Allgemeine Sprachwissenschaft und Computerlinguistik
Technische Universität Dortmund
Fakultät Informatik
Lehrstuhl für Künstliche Intelligenz

Technischer Bericht

Nr. 2016/2 (Meilenstein 4b)

Integration der KobRA-Verfahren in WebLicht

BMBF-Verbundprojekt:

Korpus-basierte linguistische Recherche und Analyse mithilfe von Data-Mining (KobRA)

Förderkennzeichen: 01UG1245C
Projektlaufzeit: 01.09.2012 bis 31.12.2015
Bearbeiter/innen: Marie Hinrichs

Tübingen, den 24.2.2016

Das diesem Bericht zugrunde liegende Vorhaben wird mit Mitteln des Bundesministeriums für Bildung und Forschung (BMBF) unter den Förderkennzeichen 01UG1245A-D gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Autoren.

Integration der KobRA-Verfahren in WebLicht

1. WebLicht Beschreibung
2. Integration von KobRA-Verfahren in WebLicht
3. Zitierte Literatur

1. WebLicht Beschreibung

WebLicht (Web-based Linguistic Chaining Tool) (Hinrichs et al., 2010) ist eine virtuelle Forschungsumgebung, in der Nutzer Verarbeitungsketten zur linguistischen Annotation erstellen und ausführen können. Das vorrangige Ziel von WebLicht ist es, Forschenden in den Geistes- und Sozialwissenschaften einfachen Zugang zu einer breiten Auswahl von Textverarbeitungswerkzeugen zu gewähren. WebLicht ist Teil der CLARIN-Infrastruktur (Dima et al., 2012) und benutzt Komponenten derselben, so z.B. die Center Registry und die CLARIN Identity Federation[1].

Weblicht basiert auf Service Oriented Architecture (SOA) (Binildas et al., 2008) Prinzipien, d.h. Verarbeitungswerkzeuge sind als Webservice implementiert, die auf Servern von verschiedenen CLARIN-Zentren im Web bereitgestellt werden.

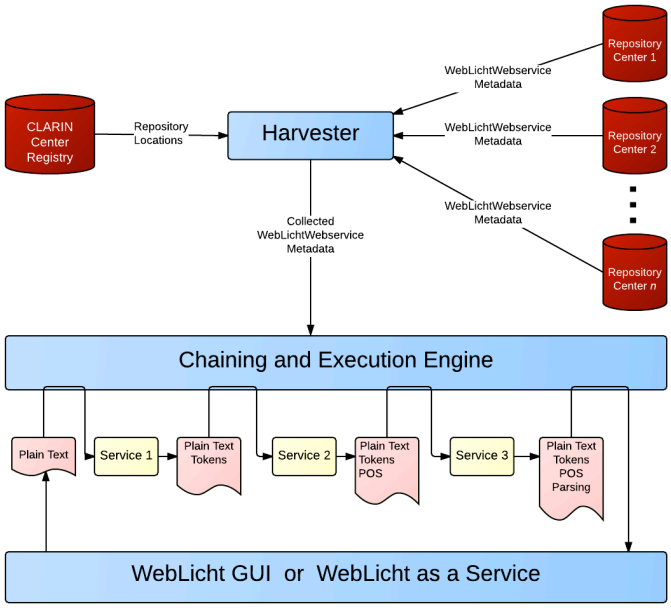

Abbildung 1: WebLicht Architektur

Die Hauptkomponenten von WebLicht und deren Interaktionen sind in Abbildung 1 zu sehen. Mit Hilfe von Informationen aus der CLARIN center registry ruft der Harvester regelmäßig Metadaten zu WebLicht-kompatiblen Webservices in den Repositorien aller Zentren ab. Die *chaining and execution engine* nutzt diese Metadaten und die Inputdaten, um Annotationsworkflows zu erstellen. Die Validität dieser Workflows wird sichergestellt, indem nach jedem

[1] http://www.clarin.eu/node/3788

Verarbeitungsschritt mögliche Services zur Weiterverarbeitung vorgeschlagen werden. Die Verarbeitungskette wird dann sequentiell ausgeführt; der Inhalt der Anfrage an Service n+1 ist das Ergebnis von Service n.

Da verschiedene Werkzeuge verschiedene In- und Output-Formate haben, wurde zum internen Datenaustausch das Text Corpus Format (TCF) (Heid et al., 2010) entwickelt. Annotationswerkzeuge stehen als Webservices, die TCF annehmen und ausgeben zur Verfügung. Die Nutzung von TCF ist keine Voraussetzung. Dennoch nutzen die meisten WebLicht-Services TCF, um mit möglichst vielen Werkzeugen kombinierbar zu sein.

Die WebLicht-Benutzeroberfläche[2] ist das Frontend zu diesen Komponenten und bietet eine einfache Möglichkeit, Workflows zu erstellen, auszuführen und die Ergebnisse zu visualisieren. Mit Hilfe der CLARIN Identity Federation können Forscher sich mit den Zugangsdaten der eigenen Universität oder des eigenen Instituts einloggen, wodurch eine breite akademische Nutzergemeinschaft Zugang erhält.

2. Integration von KobRA-Verfahren in WebLicht

Topic-Modellierung spielt eine zentrale Rolle im KobRA-Projekt. Topic-Modelle sind statistische Modelle, die ein Inputdokument in ein Set abstrakter Themen kategorisieren und mit verschiedenen Gewichten oder Prioritäten versehen. Topic-Modelle werden typischerweise automatisch aus einem Set von Dokumenten abgeleitet, ohne dass eine manuelle Annotation notwendig ist. Die resultierenden Modelle können in verschiedensten Aufgaben im Bereich des Natural Language Processing genutzt werden, z.B. zum automatischen Klassifizieren von Dokumenten oder zur Bestimmung verschiedener Wortbedeutungen. Alternativ können sie auch als "Data Mining" Tool genutzt werden, vor allem in Kombination mit passenden Visualisierungen. Auch in den digitalen Geisteswissenschaften ist Topic-Modellierung eine gängige Forschungsmethode. Insbesondere kann Topic-Modellierung dabei helfen, Muster in großen Textkollektionen zu erkennen.

Im KobRA-Projekt wurde eine Topic-Modellierungstechnik genutzt, die auf der weithin bekannten Latent Dirichlet Allocation (LDA) Technik, wie von Blei et al. (2003) beschrieben, basiert. Diese Technik wurde als Webservice in die WebLicht-Umgebung eingefügt. Nach der automatischen Erstellung eines Topic-Modells wird dieses mit der weit verbreiteten Visualisierungssoftware DFR-browser[3] abgebildet. Der DFR-Browser bietet vielfältige Visualisierungen, unter anderem von Listen der "Topwörter" zu jedem Topic und von Topic-übergreifenden Worträngen.

Bei der Topic-Modellierung sieht ein üblicher Ablauf wie folgt aus: Der Nutzer lädt eine Textsammlung in WebLicht hoch; WebLicht berechnet mit dem oben genannten Webservice ein Topic-Modell mit einer vorgegebenen Anzahl an Topics; die resultierenden Topic-Wort- und Topic-Dokument-Verteilungen werden in das vom DFR-Browser benötigte Format konvertiert und im Webbrowser des Nutzers visualisiert. Abbildung 2 enthält eine der Visualisierungsansichten und zeigt die am höchsten eingestuften Wörter für sechs abstrakte Topics in

[2] https://weblicht.sfs.uni-tuebingen.de/weblicht/

[3] Andrew Goldstone, http://agoldst.github.io/dfr-browser/

einem Topic-Modell berechnet auf Basis eines großen Zeitungskorpus. Die identifizierten Topics korrespondieren grob mit den Themenfeldern *Kultur*, *Finanzen*, *Reisen*, *Politik* und *Familie*. Solch ein Modell kann beispielsweise verwendet werden, um Artikel aus einem bestimmten Themenfeld auszuwählen bevor weitere automatische oder manuelle Analysen erfolgen.

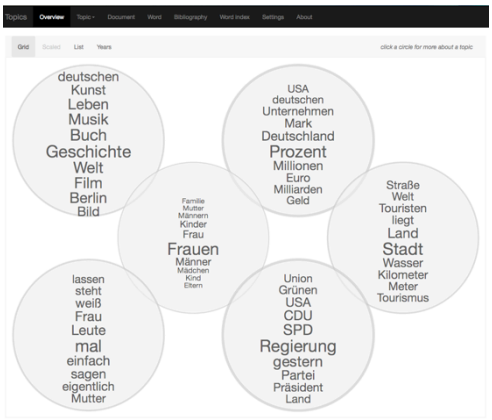

Abbildung 2: DFR-Browser Visualisierung

Der vorgestellte Webservice arbeitet zur Zeit mit mehreren Dokumenten, die als einzelne Textdatei ohne Metadaten formatiert sein müssen. Der Webservice erlaubt es derzeit, die gewünschte Anzahl von Themen einzustellen, sowie die häufigsten (zum Beispiel Funktionswörter) oder seltensten (zur Vermeidung von statistischen Störeinflüssen) Wörter herauszufiltern. Die Ergebnisse werden als statische HTML Seite dargestellt, welche auf dem Server gespeichert wird.

Dank der Zusammenarbeit mit dem KobRA-Projekt wurde ein neuer Webservice zur Topic-Modellierung in WebLicht integriert. Der LDA-Modellierungs-Service steht nun der gesamten CLARIN-Community per single-sign-on zur Verfügung.

3. Zitierte Literatur

Binildas, C.A., Barai, M. and Cas, V. (2008). *Service Oriented Architectures with Java*. PACKT Publishing, Birmingham – Mumbai.

Blei, D.M., Ng, A.Y. & Jordan, M.I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Dima, E., Hinrichs, E., Hinrichs, M., Kislev, A., Trippel, T. & Zastrow, T. (2012). Integration of WebLicht into the CLARIN Infrastructure. In *Proceedings of the joint CLARIN-D/DARIAH Workshop "Service-oriented Architectures (SOAs) for the Humanities: Solutions and Impacts" at Digital Humanities Conference 2012*, 17–23. Abgerufen unter http://clarin02.ims.uni-stuttgart.de/images/workshops/proceedingssoasforthehumanities.pdf.

Hinrichs, E., Hinrichs, M. & Zastrow, T. (2010). WebLicht: Web-based LRT services for German. In *Proceedings of the ACL 2010 System Demonstrations* (S. 25– 29). Abgerufen unter http://www.aclweb.org/anthology/P10-4005?CFID=751824403&CFTOKEN=50386258