Technischer Bericht der AG Texttechnologie

Number: TTLab-TR-2016-12

CEDIFOR-Forschungs-Workflow

AG Texttechnologie
Goethe-Universität Frankfurt am Main
Fachbereich für Informatik und Mathematik
Robert-Mayer-Straße 10
D-60325 Frankfurt am Main
Prof. Dr. Alexander Mehler

Workflow Wikidition

Ansprechpartner: Prof. Dr. Alexander Mehler

Zusammenfassung
Wikidition ermöglicht die automatische Erstellung von Online-Edition aus Textkorpora und deren Wiki-basierte Nachnutzung bzw. Erweiterung. Die Wikidition eines Korpus dient dazu, dessen semantische Struktur auf Wort-, Satz und Textebene sichtbar und nach dem Wiki-Prinzip rezipierbar bzw. erweiterbar zu machen. Darüber hinaus beinhaltet Wikidition eine *Lexiconization* genannte Komponente, welche dazu dient, das dem Inputkorpus zugrundeliegende Autorenlexikon nach dem Wiktionary-Prinzip zu extrahieren und verfügbar zu machen. Der Workflow und die involvierten Ressourcen werden beispielsweise im CEDIFOR-Projekt Historische Semantik, Kookkurrenzanalyse und Kollationierung eingesetzt.

Inhaltsverzeichnis

1. Hypothesenbildung

Am Anfang der Erstellung einer Wikidition steht die Formulierung von Forschungsfragen an Texten, welche mittels texttechnologischer Datenanalyse und Datenvisualisierung beantwortbar sein sollen.

2. Datengewinnung

Die Wikidition-Software stellt einen technologischen Rahmen zur automatischen Analyse, Aufbereitung, Darstellung und Visualisierung von Texten, Textnetzwerken, Lexika und entsprechenden Metadaten bereit. Der Nutzer kann im Prinzip Textressourcen beliebiger Art mittels Wikidition verarbeiten, wobei gemäß dem aktuellen Release von Wikidition *Deutsch*, *Englisch* und *Latein* unterstützt werden. In Arbeit sind darüber hinaus Schnittstellen für *Arabisch*, *Japanisch*, *Russisch* und *Spanisch*.

3. Datenmodellierung

Wikidition bildet ein Interface zwischen der Vielfalt computerlinguistischer Werkzeuge für die automatische Analyse natürlichsprachlicher Texte einerseits und der Wiki-basierten Rezeption der entsprechenden Analyse- und Annotationsergebnisse andererseits. Um diese Vielfalt zu überbrücken bedarf es in Absprache mit den entsprechenden „Wikiditoren“ (d.h., mit den Editoren der jeweiligen Wikidition) eines Prozessmodells, anhand dessen die Analysetiefe und Annotationsbreite der Wikidition festzulegen ist. Diese Aufgabe wird im Rahmen des Arbeitsschritts *Datenmodellierung* adressiert. Ihre Erledigung ist unabdingbar für die Gewährleistung von Interpretierbarkeit in Bezug auf die automatisch zu erstellenden Textannotationen und -verlinkungen (siehe auch Abschnitt 8).

4. Datenannotation

Wikidition erfasst morphosyntaktische Informationen lexikalischer Einheiten bis hin zu textbezogenen Daten und deren Verlinkung und erlaubt darüber hinaus die automatische Extraktion eines korpusspezifischen Lexikons. Ein besonderes Augenmerk liegt auf der Vernetzung der solcherart erfassten sprachlichen Einheiten: So besteht beispielsweise die Möglichkeit, den syntagmatischen oder paradigmatischen (semantischen) Vernetzungsrelationen von Wörtern, Sätzen und Texten nachzuspüren. Dies kann entweder horizontal (unter Bezug auf dieselbe Sprachebene) oder vertikal (unter Wechsel der Sprachebene) vollzogen werden (Abbildung 1 stellt die in einer Wikidition gegebenen Vernetzungsresultate

schematisch dar). Ferner können Sätze, Lemmata oder morphologische Worteigenschaften mit einem Klick aufgerufen werden: jede Spracheinheit wird mittels einer separaten Page charakterisiert und ist über diese Seite mit allen ihren Token im jeweiligen Inputkorpus verlinkt. Auf dieser Grundlage enthält Wikidition de facto eine dem Wiktionary-Prinzip folgende Darstellung der Lexika aller im Inputkorpus verwendeten Wörter – auf der Ebene von Lemmata ebenso wie auf der Ebene von syntaktischen Wörtern und Wortformen. Zwecks NLP und automatischer sprachlicher Vorverarbeitung setzt Wikidition auf dem TextImager auf. Auf dieser Grundlage stehen unter anderem die Tokenisierung, die Lemmatisierung, die Satzsegmentierung, das *Part-of-Speech*-Tagging, die *Named Entity Recognition*, die Erkennung von Zeitausdrücken, das Dependenzparsing, das *Semantic Role Labeling*, die Anaphernresolution und das Wikifying als grundlegende Vorverarbeitungsschritte für die automatische Datenannotation und die Berechnung syntagmatischer und paradigmatischer Vernetzungsrelationen zur Verfügung.

Darüber hinaus kann jede Wikidition nach dem Wiki-Prinzip von ihren autorisierten Nutzerinnen und Nutzern weiterverarbeitet, korrigiert oder ergänzt werden. Auf diese Weise verschränkt Wikidition das NLP und Machine Learning mit dem Human Computation. Eine Prozessübersicht diese Arbeitsschritte der Erstellung einer Wikidition gibt Abbildung 2.

5. Datenanalyse

Wikidition unterstützt eine Vielzahl an Datenexplorationen. Zur Unterstützung der geisteswissenschaftlichen Arbeit in der Historischen Semantik etwa stellt Wikidition KWIC-Ansichten bereit, deren Belegstellen allesamt mit den jeweiligen Wort- und Satz-Types verlinkt sind. Auf diese Weise können auch KWIC-Ansichten dazu verwendet werden, das wikifizierte Korpus syntagmatisch oder paradigmatisch zu traversieren. Ferner werden Graph-Ansichten für einzelne Vernetzungsrelationen angeboten: diese graphischen Ansichten enthalten Wort- oder Satzknoten, welche ihrerseits mit den Ziel-Types verlinkt sind. Auf diese Weise ist das wikifizierte Korpus nicht nur rein symbolisch, sondern auch diagrammatisch rezipierbar. Schließlich publiziert eine Wikidition sämtliche numerischen ML-Ergebnisse nach dem Prinzip der informationellen Nachhaltigkeit, welche für die Zeichenvernetzung erzeugt wurden, so dass diese Repräsentationen unabhängig von Wikidition nachgenutzt werden können.

6. Metadaten/Verlinkung

Wikidition ermöglicht die Erfassung und Repräsentation der jeweiligen Primärdaten, der hieraus gewonnenen Sekundärdaten in Form computerlinguistischer Annotationen wie auch der nutzerseitig hinzugefügten Metadaten. In Wikidition können je nach Bedarf und Zugriffsrechten textuelle Daten einem geschlossenen Nutzerkreis oder einer größeren Forschungsgemeinschaft zugänglich gemacht werden. Prospektiv beinhaltet eine derartige Veröffentlichung auch eine Bereitstellung für das *Clarin Virtual Language Observatory*, für welches bereits eine Anbindung implementiert wurde.

7. Evaluation

Die von Wikidition genutzten Ressourcen und Werkzeuge sind allesamt frei verfügbar und unabhängig von Wikidition evaluiert worden. Aufgrund des Wiki-Prinzips, das der Rezeption einer Wikidition zugrunde liegt, können alle Ergebnisse der Wikidition nutzerseitig nachgeprüft und bewertet werden. Diese Bewertungen können in Rücksprache mit dem Text-technology Lab dazu herangezogen werden, zukünftige Wikiditionen zu verbessern oder die gegebene Wikidition einer Revision zu unterziehen.

8. Rückfragen und Kontakt

Unabhängig von der technologischen Absicherung des Wikidition-Workflows steht das Text-technology Lab ggf. bei der Durchführung der einzelnen Arbeitsschritte beratend zur Seite. Es begleitet Kooperationspartner durch den Prozess der Erstellung einer eigenen Wikidition ihrer Texte und Textkorpora und unterstützt zudem die Verwaltung und Nachnutzung der erzeugten Wikiditionen.

Literatur

Mehler, Alexander, Rüdiger Gleim u. a. (2016). „Wikidition: Automatic Lexiconization and Linkification of Text Corpora“. In: *Information Technology*, S. 70–79. DOI: http://dx.doi.org/10.1515/itit-2015-0035.

Mehler, Alexander, Benno Wagner und Rüdiger Gleim (2016). „Wikidition: Towards A Multi-layer Network Model of Intertextuality“. In: *Proceedings of DH 2016, 12-16 July*. DH 2016. Accepted. Kraków.

Wagner, Benno, Alexander Mehler und Hanno Biber (2016). „Transbiblionome Daten in der Literaturwissenschaft. Texttechnologische Erschließung und digitale Visualisierung intertextueller Beziehungen digitaler Korpora“. In: *DHd 2016*. Accepted.

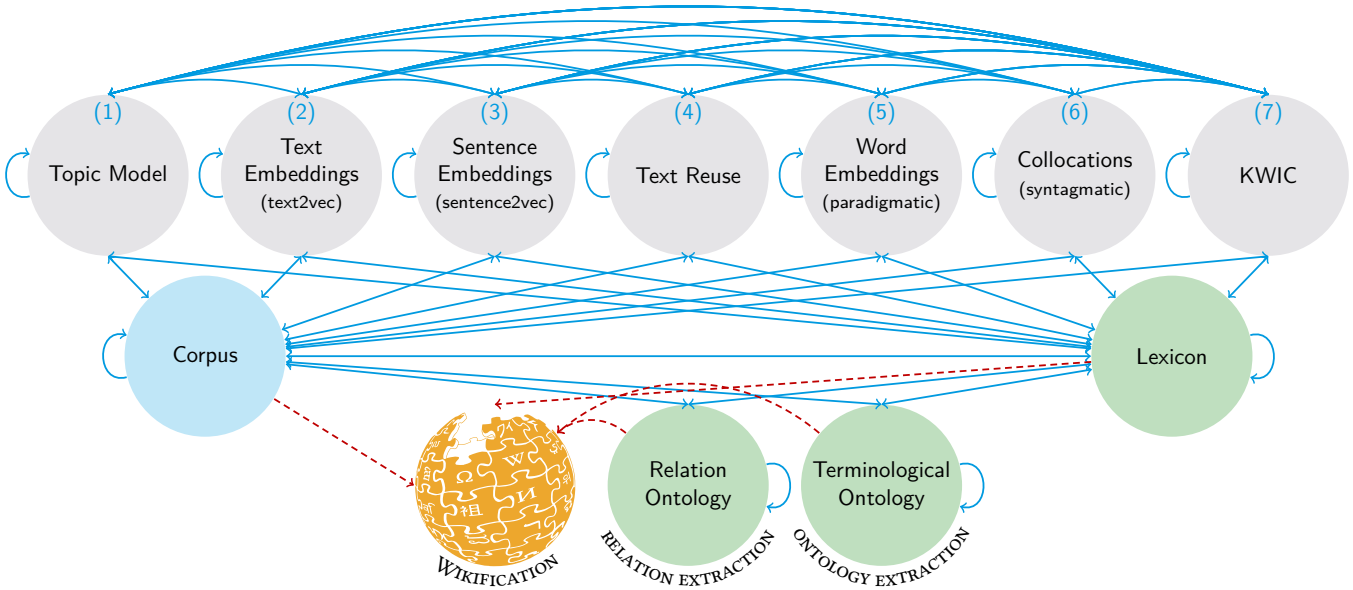

Abbildung 1. Sprachliche Morphismen zwischen den Informationseinheiten einer Wikidition (die Abbildung ist Mehler, Gleim u. a. 2016 entnommen).

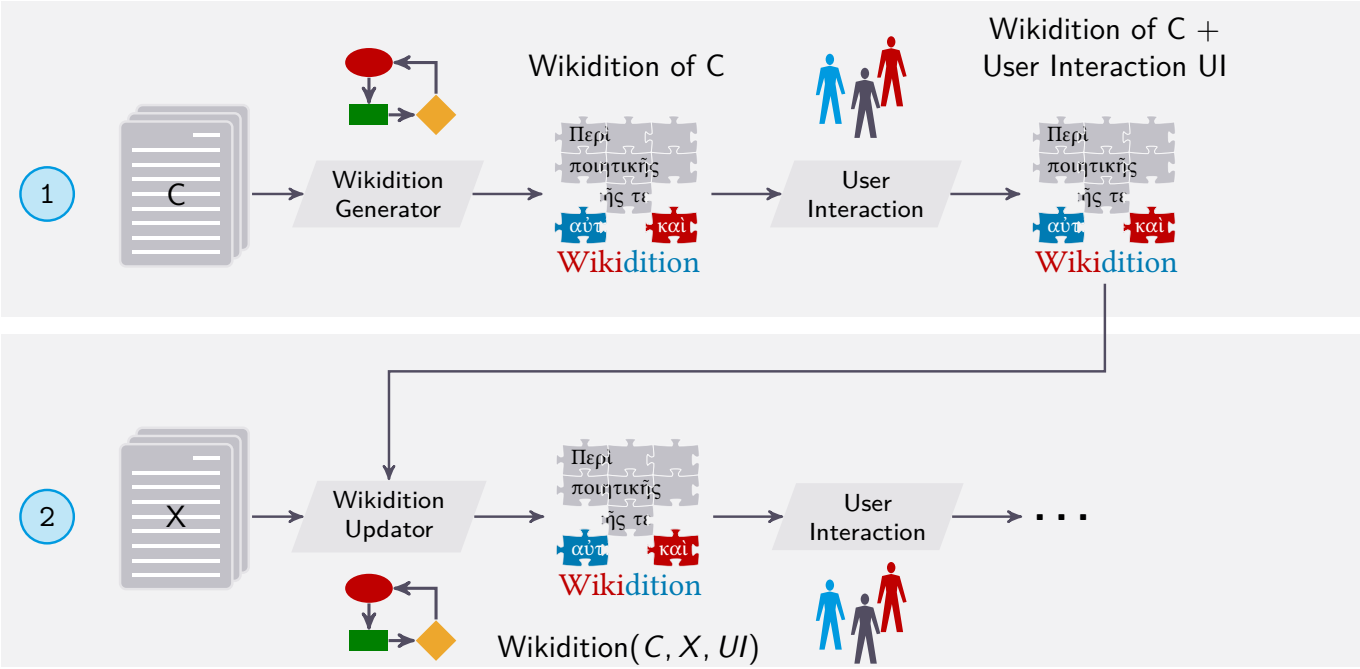

Abbildung 2. Visuelle Veranschaulichung der Prozesskette von Wikidition (vgl. Mehler, Gleim u. a. 2016).