DARIAH-DKPro-Wrapper

,
reimers@ukp.informatik.tu-darmstadt.de
TU Darmstadt, Deutschland

,
fotis.jannidis@uni-wuerzburg.de
Universität Würzburg, Deutschland

,
pielstroem@biozentrum.uni-wuerzburg.de
Universität Würzburg, Deutschland

,
stefan.pernes@uni-wuerzburg.de
Universität Würzburg, Deutschland

,
isabella.reger@uni-wuerzburg.de
Universität Würzburg, Deutschland

Dieses Poster soll den DARIAH-DKPro-Wrapper vorstellen, der aus einer Kooperation zwischen dem Lehrstuhl für Computerphilologie der Universität Würzburg und dem Ubiquituous Knowledge Processing Lab der TU Darmstadt im Rahmen von DARIAH-DE entstanden ist.

DKPro integriert zahlreiche (unabhängig entstandene) Softwarekomponenten zum Natural Language Processing (NLP) und ermöglicht so dem Nutzer die Anwendung typischer NLP-Aufgaben wie Tokenisierung, Part-of-Speech-Tagging, Named Entity Recognition oder Dependency Parsing mit State-of-the-Art Werkzeugen. Es basiert auf dem Framework UIMA. Für Nutzer, die nicht aus dem Umfeld der Informatik oder Computerlinguistik kommen, ist die Schwelle zur Verwendung allerdings recht hoch: das komplexe Framework muss in Java angesprochen werden.

Um diese Hürde zu senken und einer größeren Zahl auch von weniger technisch versierten Nutzern die Verwendung zu ermöglichen, wurde der DARIAH-DKPro-Wrapper entwickelt. Dieser ermöglicht es, eine Pipeline mit mehreren Komponenten über die Kommandozeile auszuführen und damit auch längere Textdokumente und Textsammlungen zu verarbeiten. Zudem können eine ganze Reihe von Einstellungen bequem und individuell über Konfigurationsdateien vorgenommen werden: über die Auswahl der Sprache bis hin zur Aktivierung und Deaktivierung einzelner Komponenten und der Auswahl bestimmter Komponenten oder Modelle. Auf diese Weise kann jeder Nutzer vorgefertigte Pipelines verwenden oder eine auf seine Bedürfnisse zugeschnittene Pipeline individuell zusammenstellen. Der Wrapper ist stets aktuell über GitHub (https://github.com/DARIAH-DE/DARIAH-DKPro-Wrapper) verfügbar, ebenso wie die dazugehörige Dokumentation des DARIAH-DKPro-Wrapper v0.4.3 (2016).

Um die anschließende Weiterverarbeitung derart prozessierter Dokumente ebenfalls zu vereinfachen, wurde ein entsprechendes Ausgabeformat entwickelt. Dieses lehnt sich an das CoNLL2009-Format an und stellt die Ergebnisse der Pipeline in tabellarischer Form dar. Dabei befindet sich in jeder Zeile ein Token, während die dazugehörigen Informationen wie Lemma, POS-Tag und ähnliches je in einer Spalte stehen. Dadurch werden alle durch Komponenten der Pipeline ermittelten Informationen in einer Datei zusammengefasst. Dieses Format hat den Vorteil, dass es für menschliche Nutzer übersichtlich und gut lesbar ist. Zudem ist es als Tabstopp-getrennte Datei auch für gängige Skriptsprachen wie Python oder R, sowie Tabellenkalkulationsprogramme wie Excel leicht zugänglich.

Um die Verwendung des Wrappers und die Weiterarbeit mit dem Ausgabeformat zusätzlich zur Dokumentation anschaulich zu beschreiben, wurden außerdem eine Reihe von Tutorials zu Beispielanwendungen aus Bereichen der digitalen Literaturwissenschaft, wie zum Beispiel der Stilometrie oder dem Topic Modeling, verfasst. Die Dokumentation sowie die Tutorials sind ebenfalls auf GitHub zu finden.

Das Poster wird all diese Punkte in übersichtlicher Form zusammenführen und potentiellen Nutzern präsentieren. Dabei werden die Funktionsweise der Pipeline, die Arbeit mit den Konfigurationsdateien, der Aufbau und die Verwendung des Ausgabeformats sowie Anwendungsbeispiele im Mittelpunkt stehen.

Bibliographie

Dokumentation: DARIAH-DKPro-Wrapper v0.4.3 (2016): *User guide DARIAH-DKPro-Wrapper v0.4.3* DARIAH2 - Cluster 5, Use Case 1 Team. Universität Würzburg, TU Darmstadt - DARIAH-DE https://rawgit.com/DARIAH-DE/DARIAH-DKPro-Wrapper/master/doc/user-guide.html [letzter Zugriff 08. Januar 2016].

CoNLL-2009 Format (2008-*): *CoNLL-2009 Shared Task*. Syntactic and Semantic Dependencies in Multiple Languages. Institute of Formal and Applied Linguistics, Charles University in Prague, Czech Republic, Faculty of Mathematics and Physics https://ufal.mff.cuni.cz/conll2009-st/task-description.html [letzter Zugriff 08. Januar 2016].