Ubiquitous Knowledge Processing Lab
Prof. Dr. Iryna Gurevych
Technische Universität Darmstadt
Hochschulstraße 10, 64289 Darmstadt
https://www.ukp.tu-darmstadt.de

Forschungsworkflow zur Automatischen Annotation von Textuellen Daten mit DKPro TC

Ansprechpartner: Prof. Dr. Iryna Gurevych, Dr. Johannes Daxenberger

Einsatzzweck: Automatische Identifikation von beliebigen Kategorien in textuellen Daten, basierend auf beliebig komplexen Merkmalen (bspw. lexikalische, grammatikalische, semantische etc.) und Beispieldaten (annotierte Trainingsdaten).
Fachdisziplin: alle
Anwendungsbeispiel: Automatisches Auffinden von Argumentationsstrukturen in transkribierten Interviews
Dokumentation: DKPro TC auf GitHub
Code: Quellcode DKPro TC auf GitHub
Einsatz in folgendem Assoziierten CEDIFOR Projekt: Automatisierte Analyse von wissenschaftlicher Argumentation

1. Hypothesenbildung

Hier gilt es zunächst, eine Forschungsfrage oder Hypothese so zu formulieren, dass diese anhand eines sprachlichen Phänomens untersucht werden kann.

Beispiel: Es gibt Muster wissenschaftlicher Argumentation und Schlussfolgerung, die sich in allen Berufsfeldern finden lassen.

Best Practice: Da die "Übersetzung" der Forschungsfrage/Hypothese in ein sprachliches Muster nicht trivial ist, sollte dieser Schritt im Rahmen eines Mini-Workshops zwischen einem oder mehreren Experten aus der Disziplin aus der die Forschungsfrage stammt (i.d.R. Geistes-, Sozial-, oder Bildungswissenschaft) und einem oder mehreren Informatik-Experten erörtert werden.

2. Datengewinnung

Dieser Schritt erfordert die Identifikation einer Datengrundlage, die zur Untersuchung des sprachlichen Phänomens, auf das sich in Schritt 1 geeinigt wurde, geeignet ist.

Beispiel: Transkribierte Think-Aloud Protokolle von Interviews, in denen Personen aus verschiedenen Berufsfeldern aufgefordert wurden, einen Problemfall zu erörtern.

Best Practice: Idealerweise sollte dieser Schritt zusammen mit Schritt 1 im Rahmen eines Mini-Workshops erörtert werden. Häufig ist die Datengrundlage auch bereits vorhanden (bspw. aus vorangegangenen qualitativen Untersuchungen).

3. Datenannotation

Sobald die Datengrundlage bekannt ist, müssen die sprachlichen Phänomene, die untersucht werden sollen (Schritt 1), anhand eines klar definierten Modells (mit finiten Kategorien) kodiert werden. Das Ergebnis ist ein sogenannter "Gold Standard".

Beispiel: Zunächst wird eine Teilmenge der transkribierten Think-Aloud Protokolle für die Annotation ausgewählt. In diesen Protokollen wird jede äußerung einer oder mehreren (finiten und zuvor definierten) Kategorie(n) zugewiesen. Die Kategorien benennen bestimmte Aspekte wissenschaftlicher Argumentation (bspw. Problemerörterung).

Best Practice: Falls solche Daten nicht bereits aus vorangegangenen qualitativen Untersuchungen vorhanden sind, sollte für diesen Schritt ein geeignetes Annotationswerkzeug

verwendet werden. Wir empfehlen WebAnno. Wichtig ist, dass vor Beginn des Annotationsprozesses klare Richtlinien (Annotation-Handbuch) geschrieben werden und deren Anwendbarkeit im Rahmen einer Messung des Agreements zwischen mehreren Annotatoren verifiziert wird. Ebenso wichtig ist, dass eine Mindestmenge an zu annotierenden Daten festgelegt wird, so dass ausreichend Trainingsdaten generiert werden.

4. Experimentelle Aufbereitung/Exploration

Hier müssen zunächst geeignete Merkmale (Features), die zu untersuchende sprachliche Phänomen auf einer abstrakten Ebene beschreiben, definiert werden. DKPro TC unterstützt diesen Schritt, indem viele grundlegende Features vordefiniert sind. Wenn der finale Feature-Satz feststeht, wird ein geeignetes statisches Verfahren darauf angewandt.

Beispiel: Lexikalische und grammatikalische Features (bspw. Ngrams und Part-of-speech Tags) werden für jede Äußerung in Think-Aloud Protokollen extrahiert. Diese Features sind der Input für eine Regressionsanalyse, die versucht, die extrahierten Merkmale möglichst genau auf die in Schritt 3 definierten Kategorie(n) abzubilden.

Best Practice: Dieser Schritt erfordert Programmierkenntnisse und grundlegende Erfahrung mit statistischen Machine Learning Verfahren (Parametrisierung) und sollte von einem Informatik-Experten ausgeführt werden.

5. Datenanalyse

Das in Schritt 4 definierte Verfahren wird hier auf den annotierten Trainingsdaten angelernt. Anschließend kann es zur automatischen Annotation beliebig großer Daten (siehe Schritt 2 eingesetzt werden.

Beispiel: Das trainierte Modell wird zur automatischen Annotation der nicht annotierten Protokolle herangezogen. Anschließend werden die Kategorien wissenschaftlicher Argumentation, getrennt nach Berufsgruppe, ausgewertet (bspw. Verteilung der Kategorien oder deren zeitlicher Verlauf). Diese Auswertungen können für jede Berufsgruppe verglichen werden, um eventuelle allgemeingültige Muster wissenschaftlicher Argumentation und Schlussfolgerung zu identifizieren.

Best Practice: Die Datengrundlage, auf die das trainierte Modell angewandt wird, hängt stark von der Forschungsfrage/Hypothese (Schritt 1) ab und muss von Anfang an feststehen. Idealerweise wird in Schritt 2 ein großes Korpus ausgewählt, von dem in Schritt 3 nur eine Teilmenge annotiert wird. Der nicht-annotierte Rest kann dann automatisch annotiert werden. Auf dem Gesamtkorpus erfolgt dann die Auswertung der Forschungsfrage/Hypothese.

6. Evaluation

Die annotierten Daten werden in Trainings- und Test geteilt. Auf den Testdaten wird das in Schritt 4 gelernte Modell evaluiert (die Anzahl korrekter oder falscher Aussagen des Modells wird gemessen).

Beispiel: 75% der Aussagen in den Protokollen werden vom Modell mit korrekten Kategorien versehen.

Best Practice: Dieser Schritt sollte zusammen mit Schritt 4 erfolgen, um sicherzustellen, dass die Aussagekraft des Modells hoch genug ist, um weitere Schlüsse auf den vom Modell erstellen automatischen Annotation zuzulassen. Die Beurteilung der Ergebnisse sollte in Abgleich mit dem (relevanten) State-of-the-art erfolgen.

7. Datensicherung

Die in Schritt 2, 3 sowie 5 erzeugten Daten (“Rohdaten” mit manuell und automatisch erzeugten Annotationen) annotierten, getrennt nach “Gold Standard” und automatisch annotierten Daten, müssen in einem geeigneten Format und Ort gespeichert werden, so dass sichergestellt ist, dass die Daten für möglichst lange Zeit nachnutzbar sind.

Beispiel: Alle transkribierte Think-Aloud Protokolle werden, getrennt nach “Gold Standard” und automatisch annotierten Protokollen, in einem zum Apache-Standard UIMA kompatiblen Format gespeichert.

Best Practice: Je nach Lizenz und Verwertungsstrategie sollten die Daten entweder frei zugänglich (bspw. auf der Homepage der Beteiligten) oder intern einem Forschungsdatenmanagementsystem abgelegt werden. Als Format kommt alles in Frage, was standardisiert ist und in der Lage textuelle Daten und Metadaten (Annotationen) abzubilden (z.B. UIMA Cas, TEI-XML).