Worteinbettung als semantisches Feature in der argumentativen Analyse

Bachelorverteidigung

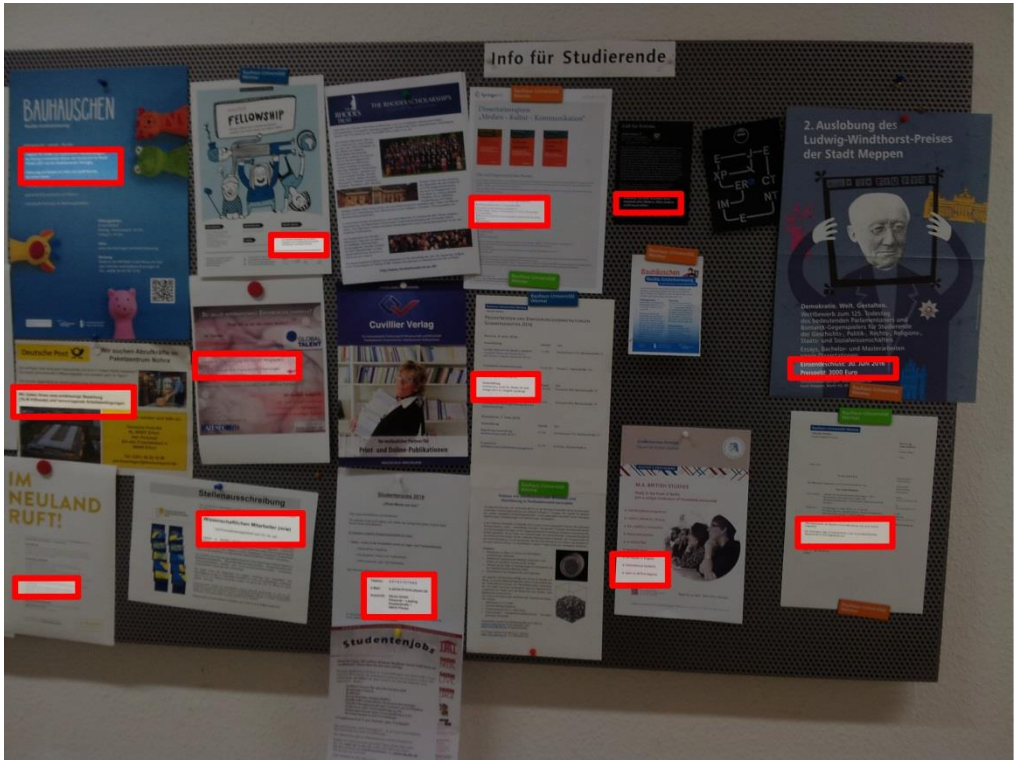
Kevin Lang

22.03.2016

Übersicht

- Was ist die argumentative Analyse?
- Worteinbettung und Word2Vec
- Resultate
- Diskussion
- Zukunftsaussichten

Motivation



- Foren
- News
- Netzwerke
- wissenschaftliche Arbeiten
- …

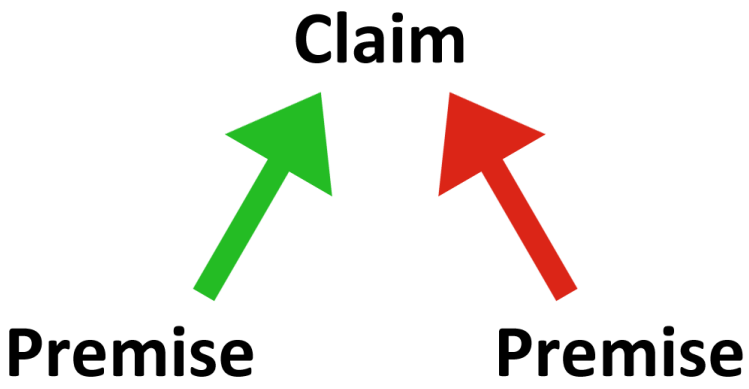
Argumentative Kernaussagen

Motivationen

- Kernaussagen eines argumentativen Textes schnell finden
- Für und Wider für aktuelle Themen auflisten
- Aussagen überprüfen und vergleichen
- Schreibunterstützung für zukünftige Texte

Argumentation

- **Claim**: Behauptung, Standpunkt des Verfassers
- **Premise**: Unterstützende oder angreifende Aussagen
- 1:n-Beziehung zwischen Claim und Premise



Die vier Schritte der argumentativen Analyse

Beispiel:

The winner is the athlete but the success belongs to the whole team. Therefore without the cooperation, there would be no victory of competition.

Die vier Schritte der argumentativen Analyse

- **Segmentierung**

Beispiel:

The winner is the athlete | but | the success belongs to the whole team. | Therefore | without the cooperation, there would be no victory of competition. |

Segmentgrenzen

Die vier Schritte der argumentativen Analyse

- **Segmentierung**
- **einfache Klassifizierung**

Beispiel:

The winner is the athlete | but [the success belongs to the whole team.] Therefore [without the cooperation, there would be no victory of competition.]

argumentative Segmente

Die vier Schritte der argumentativen Analyse

- **Segmentierung**
- **einfache Klassifizierung**
- **einfache Relationen**

Beispiel:

The winner is the athlete but the success belongs to the whole team. Therefore without the cooperation, there would be no victory of competition.

Premise → **Claim**

Die vier Schritte der argumentativen Analyse

- **Segmentierung**
- **einfache Klassifizierung**
- **einfache Relationen**
- **spezifischere Relationen**

Beispiel:

The winner is the athlete but the success belongs to the whole team. Therefore without the cooperation, there would be no victory of competition.

Premise → **Claim**

Bisherige Methoden

- Strukturelle Features
- Lexikalische Features
- Syntaktische Features
- Semantische Features

Ein neues semantisches Feature

- Worteinbettung mit Word2Vec
- Wörter haben eine bestimmte Bedeutung, die durch Vektoren ausgedrückt werden

➢ **Ziel**: Ähnlichkeiten zwischen Wörtern/Wortgruppen sollen Auskunft über Argumentation geben

Beispiel:

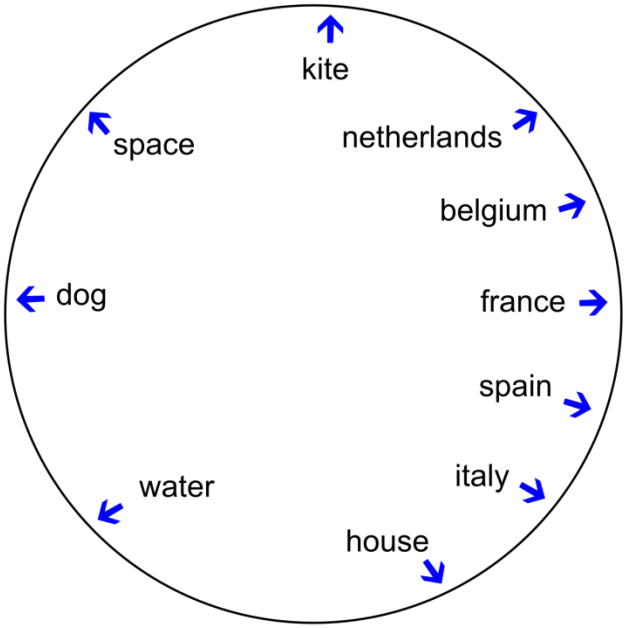
vector('king') - vector('man') + vector('woman') ≈ vector('queen')

Funktionsweise von Word2Vec

Input:
Text

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur

Modell:



Funktionen:

Bsp. most_similar('france')

spain	0.678515
belgium	0.665923
netherlands	0.652428
italy	0.633130
...	

Versuchsaufbau

Trainingsdaten und daraus entstandene Modelle:

uppsala	text8	wiki14	news
1.500 arg. Aufsätze, ca. 12MB	Wikipedia Korpus von 2006, ca. 100MB	Wikipedia Korpus von 2014, ca. 6GB	Fertiges Modell aus Google News, ~100GB

Klassifikationsdaten zum Evaluieren:

AAEC	AMC
Besteht aus 90 annotierten Schulaufsätzen mit 1552 arg. Segmenten und 1473 Relationen	Korpus in 2-Satz-Struktur zum Thema Politik mit 2274 Relationen von News Webseiten gesammelt

Resultate der Experimente

- **Experiment 1**: Verbesserung der Segmentierung

➢ Ergebnis: Bag-of-Words konnte nicht geschlagen werden, Problem der Stoppwörter

The winner is the athlete | but | the success belongs to the whole team. | Therefore | without the cooperation, | there would be no victory of competition. |

- **Experiment 2**: Klassifizierung von Segmenten

➢ Ergebnis: Bag-of-Words konnte ebenfalls nicht geschlagen werden, keine sinnvolle Berechnung von Ähnlichkeiten möglich

The winner is the athlete | but | the success belongs to the whole team. | Therefore | without the cooperation, | there would be no victory of competition. |

Resultate der Experimente

- **Experiment 3**: Relationenfindung von Argumenten

The winner is the athlete but the success belongs to the whole team. Therefore without the cooperation, there would be no victory of competition.

Premise → Claim

Klassifikation	uppsala	text8	wiki14	news
AAEC	57,86%	57,65%	57,63%	**58,77%**
AMC	59,77%	59,77%	**68,71%**	60,02%

- Verschiedene Verfahren wurden angewendet: z.B. Extrahierung von Nomen, Nominalphrasen
- Bag of Words Genauigkeit in AAEC von **58,76%** und in AMC von **66,08%** konnten knapp geschlagen werden

➢ Besten Ergebnisse stammen meist von der **Extrahierung der Nomen**

➢ Die **Analyse ganzer Sätze** im AMC Korpus brachte bessere Ergebnisse hervor als nur von Segmenten im AAEC Korpus

Diskussion

Die Trainingsdaten

uppsala	text8	wiki14	news
argumentativ, aber geringer Wortschatz und Themenumfang	enzyklopädisch, großer Wortschatz/ Themenumfang, nicht aktuell	enzyklopädisch, großer Wortschatz/ Themenumfang	argumentativ? großer Wortschatz, wahrscheinlich nicht sehr Themenumfassend

- Es existiert kein Korpus, der gleichzeitig argumentativ ist, einen großen Wortschatz hat und möglichst viele Themen/Kontexte umfasst

Diskussion

Die Klassifikationsdaten

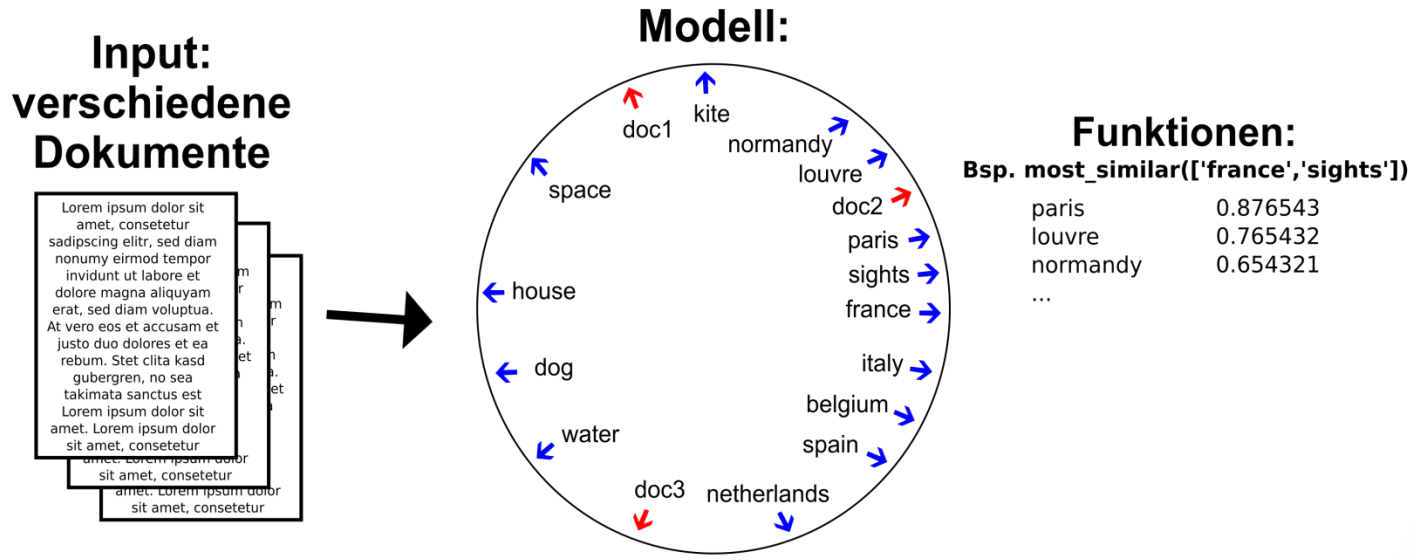
AAEC	AMC
Gute Annotierung aber Segmente geben oft nicht genug Informationen her	Ganze Sätze, dadurch mehr Informationen, aber Annotierung nicht immer nachvollziehbar

- Viel mehr support- als attack-Relationen
- Annotierung nicht immer eindeutig, Urteilübereinstimmung der annotierenden Personen liegt nur bei rund 86%

Diskussion

Das Framework Word2Vec

- Word2Vec nimmt pro Wort nur einen Kontext auf
  → Lösung: Doc2Vec



- Word2Vec oft nicht so vielversprechend wie vorgestellt

vector('king') - vector('man') + vector('woman') ≈ vector('queen')

Diskussion

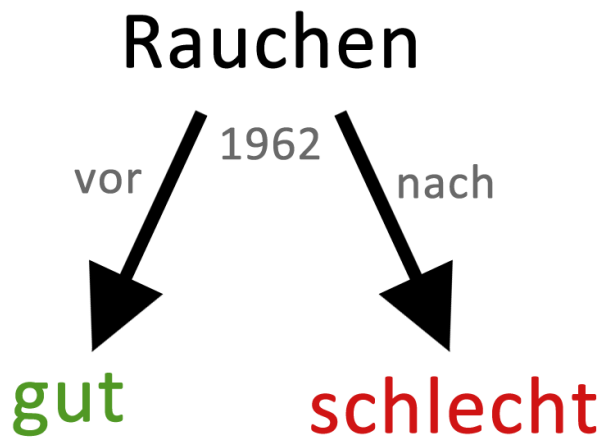
A1) vec('***library***') ⇔ vec('***book***')
A2) vec('***library***') ⇔ vec('***borrow***','***book***')
B1) vec('***productivity***') ⇔ vec('***employee***')
B2) vec('***productivity***') ⇔ vec('***hard***', '***working***', '***employee***')
B3) vec('***productivity***') ⇔ vec('***fool***', '***around***', '***employee***')
B4) vec('***productivity***') ⇔ vec('***lazy***', '***employee***')

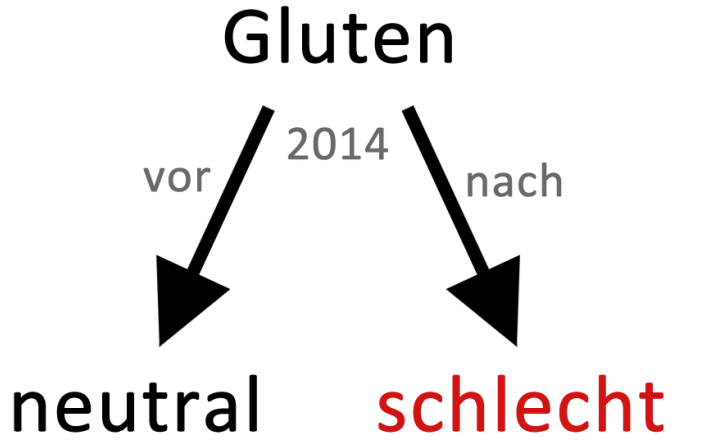
Relation	uppsala	text8	wiki14	news
A1	0,2524	0,2548	0,4871	0,3245
A2	**0,6919**	0,2387	0,4871	0,3738
B1	0,1451	0,2264	0,4416	0,2474
B2	**0,3316**	0,2868	0,4623	0,2636
B3	0,1626	0,2218	**0,3817**	0,1666
B4	0,1472	0,2320	**0,3772**	**0,3205**

Diskussion

Allgemeine Frage: Argumentationen überhaupt lernbar?

- Argumentation soll von etwas neuem überzeugen, wie dann vorher lernbar?
- Argumentationen ändern sich über die Zeit, "Gut" und "Schlecht" nicht immer klar definierbar





Zukunftsaussichten

Durch Worteinbettung mit Word2Vec können Aufgrund von Daten und funktionsweise kaum sinnvolle semantische Vorhersagen zu Argumentationen getroffen werden.

- Suche nach besseren **Trainingsdaten**, die groß, argumentativ und themenübergreifenden sind und **Klassifikationsdaten**, die verschiedene Quellen umfassen (Essays, Kommentare, News etc.)
- **Postprocessing** der Modelle durch Lexika, um die Vektoren von Wortgruppen zu verbessern
- **Doc2Vec** als Erweiterung von **Word2Vec** wenn bessere Implementationen vorhanden

Vielen Dank für Ihre Aufmerksamkeit!

Thesen

- Worteinbettungen kann die Textsegmentierung verbessern
- Durch Worteinbettungen kann man argumentative Segmente erkennen
- Durch Worteinbettungen kann man Relationen zwischen argumentativen Segmenten erkennen
- Ein größerer Trainingskorpus erzielt bessere Ergebnisse
- Das Extrahieren bestimmter Satzbestandteile führt zu besseren Ergebnissen
- Es existiert ein Grenzwert in der Kosinusähnlichkeit, ab dem man eine Relation als argumentativ einstufen kann

Resultate von Experiment 1

- **Ziel**: Verbesserung der Segmentierung

uppsala	text8	wiki14	news
56,61%	**57,68%**	55,38%	-

- Bag of Words mit **69,06%** konnte nicht geschlagen werden
- Keine ausschlagenden Ergebnisse
- Problem: meist Stoppwörter an Grenzen, die keine sinnvolle Semantik besitzen

Resultate von Experiment 2

- **Ziel**: Erkennung argumentativer Segmente

Features	uppsala	text8	wiki14	news
Word2Vec	69,59%	69,97%	**71,01%**	69,09%
Word2Vec+Pos	80,33%	80,68%	81,02%	81,13%

- Bag-of-Words Genauigkeit von **71,06%**
- Genauigkeit durch Positionsmerkmale 78,55%
- Im AAEC Korpus befinden sich meisten Argumente am Ende des Satzes (82,62%), am Ende des Paragraphen (73,59%) und in der Mitte des Essays (74,81%)

Diskussion

Das Framework Word2Vec

	uppsala	text8	wiki14	news
Ø	0.1410	0.0700	0.0819	0.1305

- Die Durchschnittsähnlichkeit von allen Modellen zeigt, dass diese oft nicht bei 0 liegt

Software

- Datenextrahierung mit den Stanford CoreNLP Tools
- Die Python Gensim Implementierung von Word2Vec
- Maschinelles Lernverfahren mit Weka