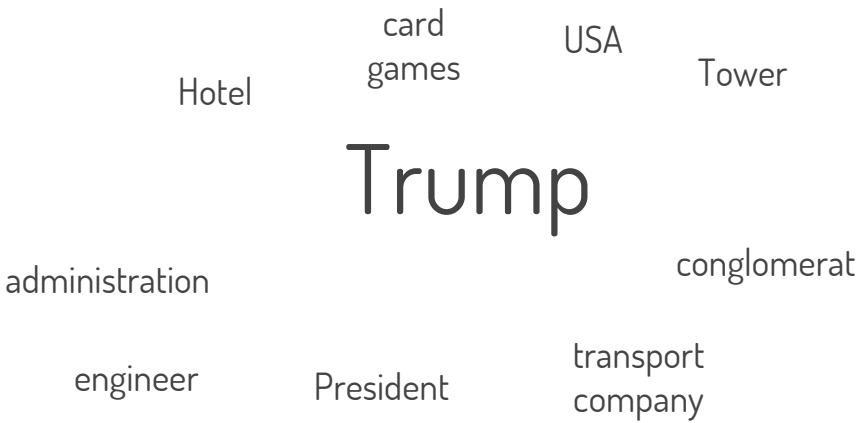
Verbessertes Entity Linking mit neuronalen Word Embeddings

Bachelorarbeit

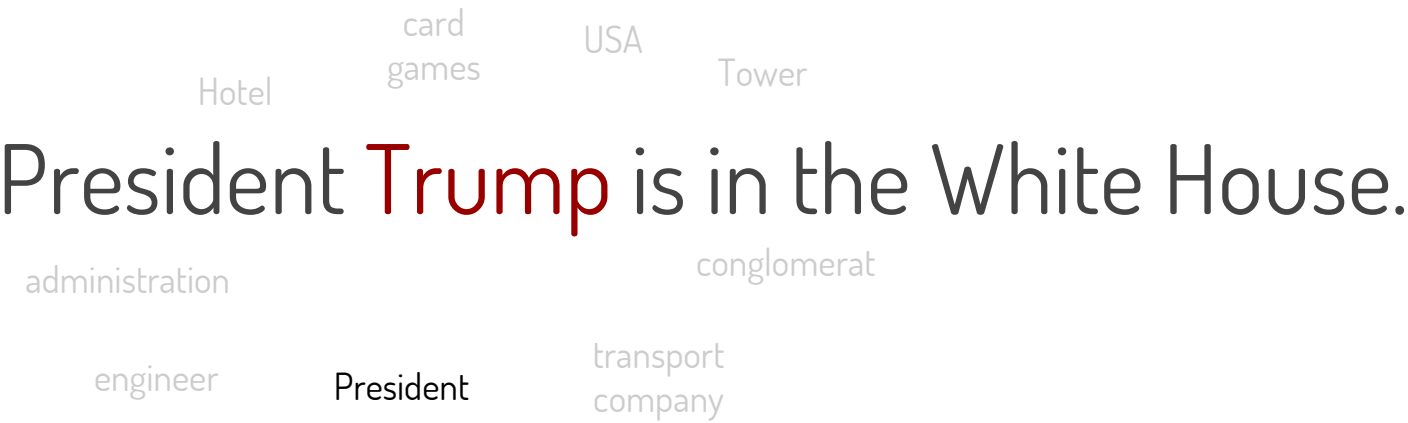
Verfasser: Robert Dziuba
Betreuer: Prof. Dr. habil. Alexander Löser
Gutachter: Prof. Dr.-Ing. Joachim Schimkat

Problemstellung

Das Wort – Mention



Der Kontext und Word Embedding



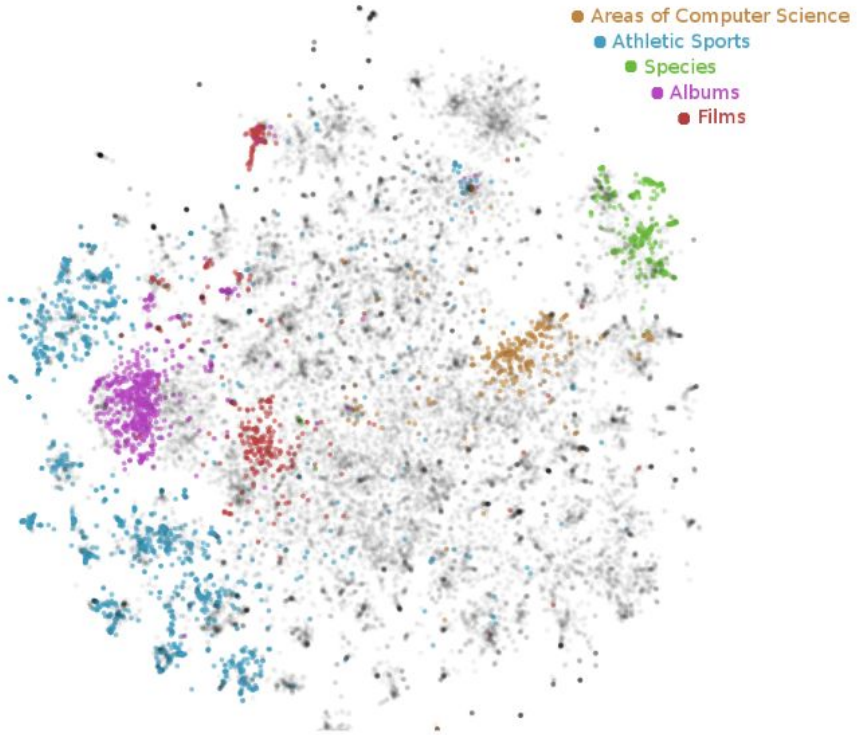
Ziel der Arbeit

Ziel der Arbeit

Das Ziel der Arbeit ist es, eine im Text genannten Mention, mit Hilfe ihres Kontextes, einer konkreten Entity zuzuordnen.

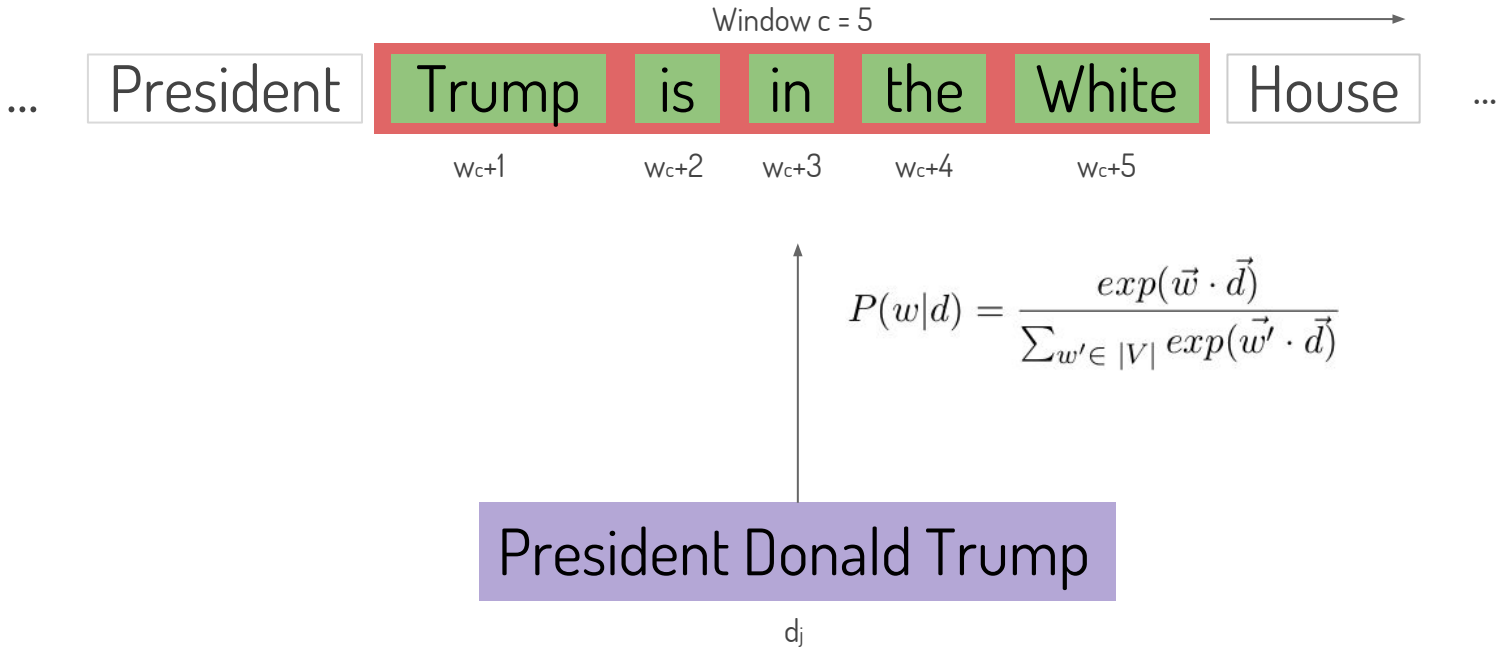
Methodik

Repräsentation von Kontext im Vektorraum - Entity Embeddings



Document Embedding with Paragraph Vectors https://arxiv.org/pdf/1507.07998.pdf

Paragraph Vector



Disambiguierung im Vektorraum

Mention = "Trump"

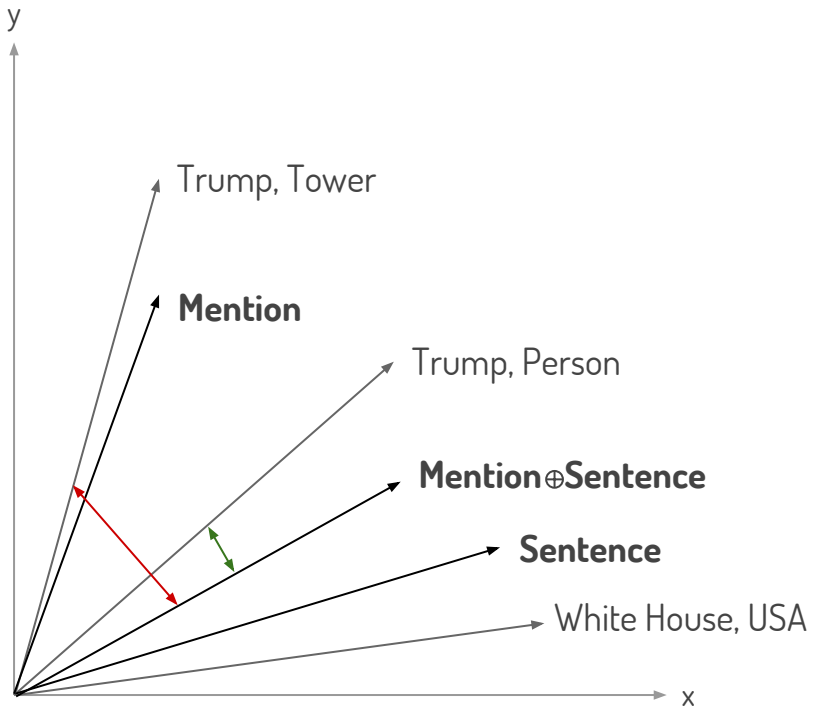
Sentence = "President Trump is in the White House."

$$cosine(\overrightarrow{v}, \overrightarrow{w}) = \frac{\overrightarrow{v} \cdot \overrightarrow{w}}{|\overrightarrow{v}||\overrightarrow{w}|} = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i=1}^{N} w_i^2}}$$

$$\hat{c} = \sum_{c \in |C_m|} cosSim(c)$$

mit $cosSim(c) = cosine(concat(m, s), concat(c, c))$

und $concat(a, b) = vec(a) \oplus vec(b)$



Daten

Wikipedia Korpus

Within American political culture, the center-right Republican Party is considered "conservative" and the center-left Democratic Party is considered "liberal".[375][376] The states of the Northeast and West Coast and some of the Great Lakes states, known as "blue states", are relatively liberal. The "red states" of the South and parts of the Great Plains and Rocky Mountains are relatively conservative.

Republican Donald Trump, the winner of the 2016 presidential election, is serving as the 45th President of the United States.[377] Leadership in the Senate includes Republican Vice President Mike Pence, Republican President Pro Tempore Orrin Hatch, Majority Leader Mitch McConnell, and Minority Leader Chuck Schumer.[378] Leadership in the House includes Speaker of the House Paul Ryan, Majority Leader Kevin McCarthy, and Minority Leader Nancy Pelosi.[379]

Donald Trump
45th President
since January 20, 2017

Mike Pence
48th Vice President
since January 20, 2017

Wikipedia: Donald Trump - https://en.wikipedia.org/wiki/United_States

Trainingsdaten

Beispielsatz

Q22686 Republican
Donald Trump, the winner of
the 2016 presidential election,
is serving as the 45th
President of the United States.

Datensatz

2.725.363 Entities

63.812.227 Sätzen

11 GB Gesamtumfang

Evaluation

Ergebnisse der Vektorkonkatenation-Strategie

Layer/Epochen	**MRR**	**PREC** (%)	**REC** (%)	**F1** (%)
Lucene	0.534	48.24	45.10	46.60
	CommonPreprocessor			
100 Layer/10 Epochen	0.592	56.73	53.10	54.84
	MinimalLowercasePreprocessor			
100 Layer/10 Epochen	0.594	57.20	53.52	55.28
	CommonPreprocessor			
300 Layer/10 Epochen	**0.630**	**62.01**	**57.96**	**59.90**

Fazit und Ausblick

Fazit

- Mention Embeddings sind besser als Lucene Exact Match
- Konkatenierte Entity Embeddings sind besser als Lucene
- Mehr Layers führen zu besseren Ergebnissen

- Trainingszeit steigt überproportional mit Iteration/ Epochen
- Lucene limitiert das Suchergebnis
- Entities die nicht in der Wikipedia verlinkt sind werden nicht gelernt

Ausblick

- Minimierung der Modellgröße
- Lucene beim Entity Linking weglassen
- Mehr Kontext
- Einfluss benachbarter Entities im Text nutzen

Live Präsentation

Vielen Dank