

Projekt Daidalos

Maschinelles Lernen für die semantische Philologie

Konstantin Schulz

Humboldt-Universität zu Berlin

July 1, 2020

Funded by

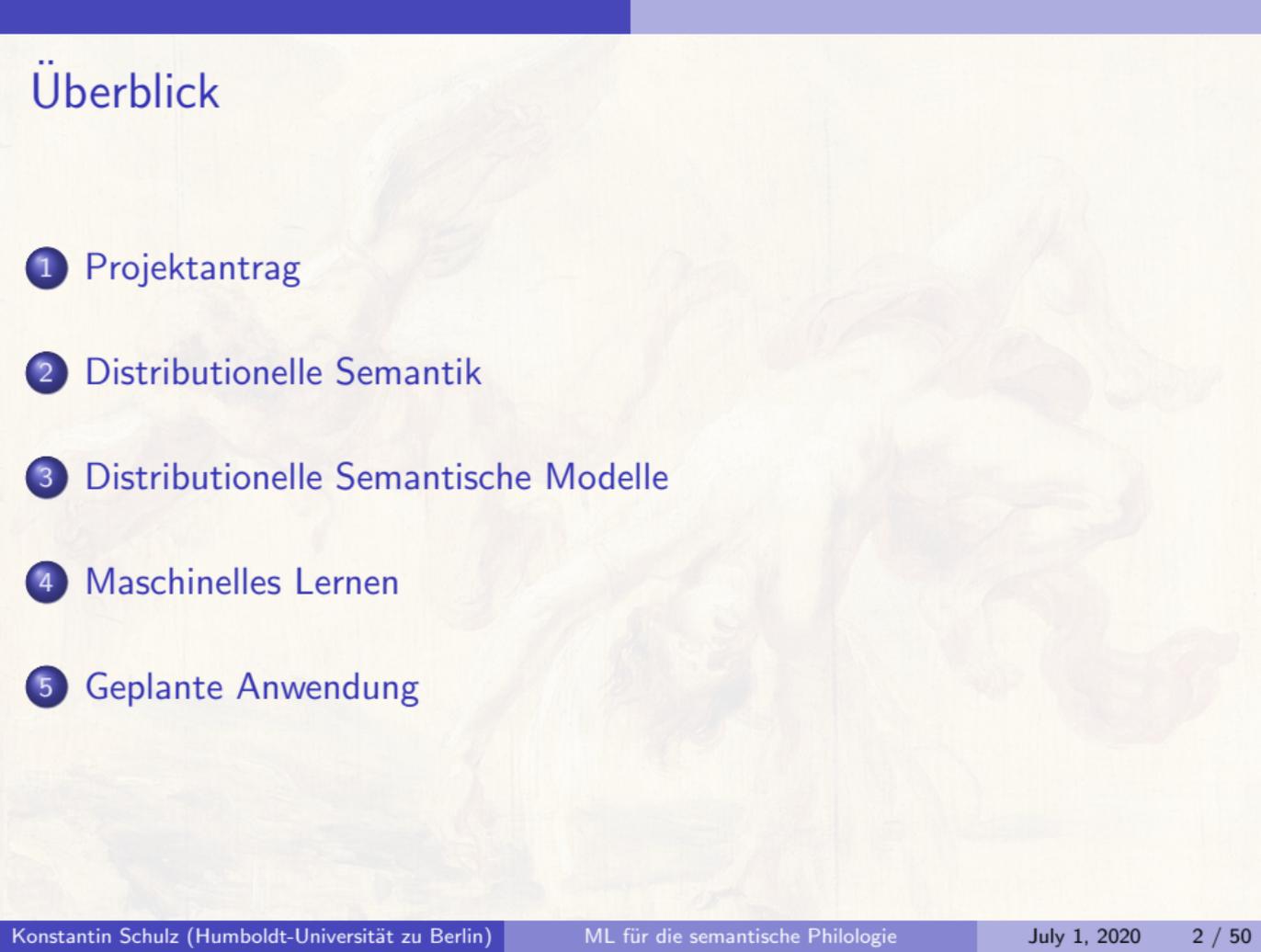
DFG

Deutsche
Forschungsgemeinschaft

German Research Foundation



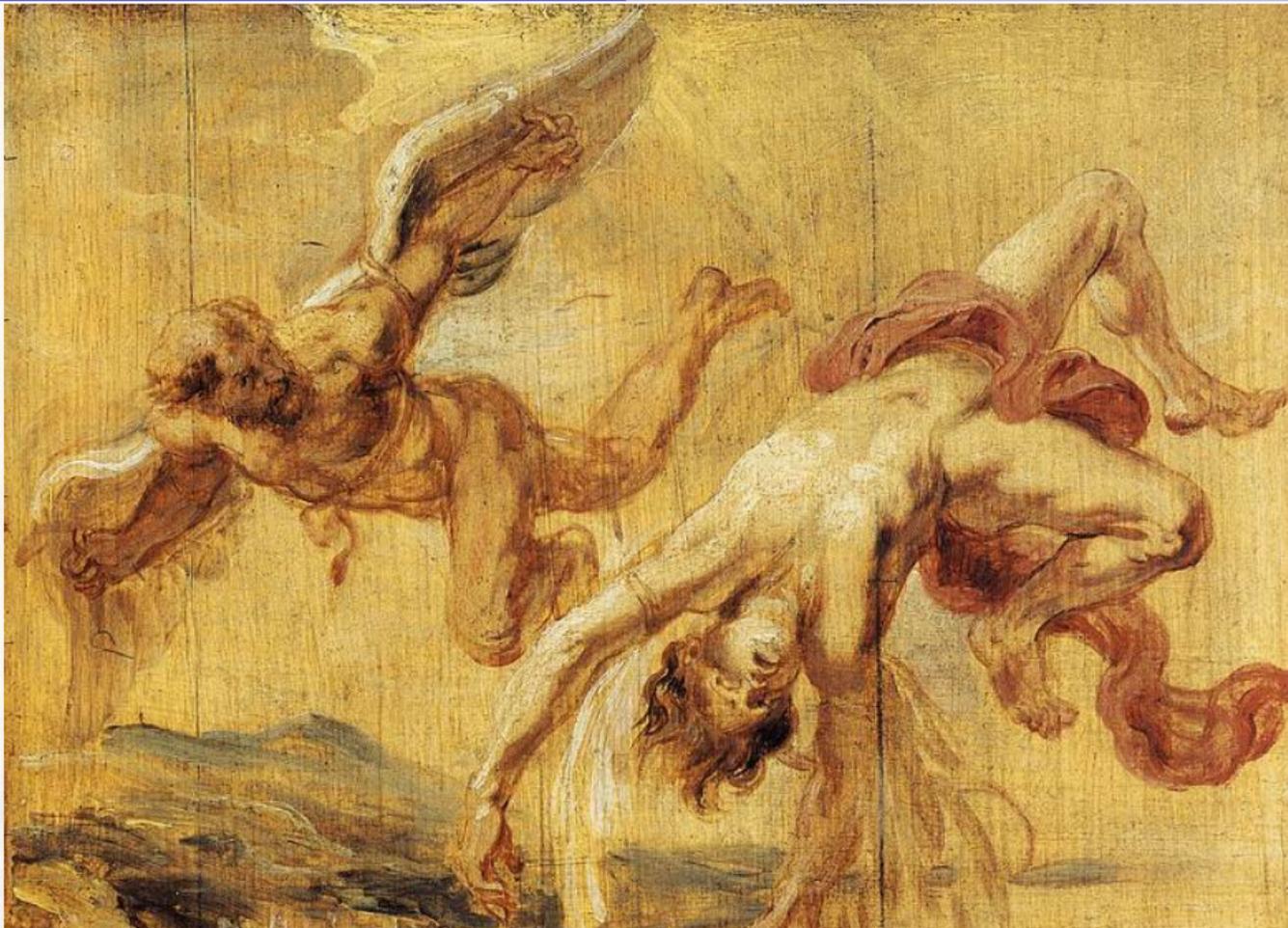
Überblick



- 1 Projektantrag
- 2 Distributionelle Semantik
- 3 Distributionelle Semantische Modelle
- 4 Maschinelles Lernen
- 5 Geplante Anwendung

The background of the slide is a faint, classical-style painting. It depicts a winged figure, possibly a cherub or putto, in the upper left, and a reclining figure, possibly a woman, in the lower right. The figures are rendered in a soft, painterly style with visible brushstrokes. The overall color palette is muted and classical, with a focus on the central text.

Projektantrag



Einordnung

- **Forschungsinfrastruktur** / e-Research-Technologie
- Ziel: philologische Forschung **digital** unterstützen
- Fokus: **semantische Suche** in literarischen Texten
- interdisziplinär: **Korpuslinguistik & CMS**, mit Anbindung an Klassische Philologie

KULTUR

Beschäftigung mit angrenzenden
Fachgebieten (z.B. Alte Geschichte,
Archäologie, Philosophie,
Byzantinistik, Lateinische Philologie,
Theologie)

SPRACHE

Übersetzen vom
Griechischen ins
Deutsche
Wiederholung und
Vertiefung der
Grammatik
Übersetzen vom
Deutschen ins
Griechische

**Altgriechische
Texte**

vom 8. Jh. v. Chr. bis
zum 6. Jh. n. Chr.

LITERATUR

Arbeit mit
Sekundärliteratur
Methoden der
Literatur-
wissenschaft
Literaturgeschichte
Rezeption

ÜBERLIEFERUNG

Editionswissenschaft/Textkritik
Paläographie und Handschriftenkunde

Konkrete Anforderungen

- relevante antike **Begriffe** zu einem Thema identifizieren, z.B. zu Fiktion und Faktualität in spätantiker Panegyrik
- intra- und **intertextuelle** Parallelen erforschen, z.B. wiederkehrende Topoi antiker Panegyrik
- autoren- oder **genrespezifischen** Sprachgebrauch untersuchen, z.B. Übertreibung in antiker Panegyrik

philologische Forschung zu antiker Panegyrik: Cordes 2020; Intertextualität: Frolov et al. 2011

Strategie

- theoretische Grundlagen
 - ▶ **distributionelle** Semantik
 - ▶ interpretierbares **maschinelles** Lernen

Explainable Artificial Intelligence: Doran et al. 2017

Strategie

- theoretische Grundlagen
 - ▶ **distributionelle** Semantik
 - ▶ interpretierbares **maschinelles** Lernen
- praktische Umsetzung
 - ▶ Evaluation: **Testdatensätze**
 - ▶ **Repräsentation**: maschinelles Lernen auf variabler Textgrundlage
 - ▶ Anwendung: Suche in **Vektorräumen**
 - ▶ Ergebnisse: interaktive **Visualisierung**

Explainable Artificial Intelligence: Doran et al. 2017



Distributionelle Semantik

Grundlagen

- **Selektion/Valenz** als Hinweis auf die Bedeutung von Wörtern
 - ▶ *princeps qui putat* (der Kaiser, der glaubt)
 - ▶ *princeps putat* (der Kaiser glaubt)

Selektion als semantisches Indiz: Harris 1954; Kollokationen und *colligations* in der distributionellen Semantik: Firth 1957; Differenzierung fast synonyme Begriffe: Lehecka 2015; Verteilung und Funktion: Gries et al. 2009

Grundlagen

- **Selektion/Valenz** als Hinweis auf die Bedeutung von Wörtern
 - ▶ *princeps qui putat* (der Kaiser, der glaubt)
 - ▶ *princeps putat* (der Kaiser glaubt)
- Analyse der Bedeutung eines Textes durch
 - ▶ Lexik, z.B. **Kollokationen**
 - ★ *patres conscripti* (werte Senatoren)
 - ★ *res publica* (Staat)

Selektion als semantisches Indiz: Harris 1954; Kollokationen und *colligations* in der distributionellen Semantik: Firth 1957; Differenzierung fast synonyme Begriffe: Lehecka 2015; Verteilung und Funktion: Gries et al. 2009

Grundlagen

- **Selektion/Valenz** als Hinweis auf die Bedeutung von Wörtern
 - ▶ *princeps qui putat* (der Kaiser, der glaubt)
 - ▶ *princeps putat* (der Kaiser glaubt)
- Analyse der Bedeutung eines Textes durch
 - ▶ Lexik, z.B. **Kollokationen**
 - ★ *patres conscripti* (werte Senatoren)
 - ★ *res publica* (Staat)
 - ▶ Syntax, z.B. **Konstruktionen** als Mittel zur Differenzierung zwischen fast synonymen Begriffen
 - ★ *imperator auguste/invicte* ([oh du] erhabener/unbesiegter 'Kaiser') vs. *princeps noster fecit* (unser 'Kaiser' machte)
 - ★ *fieri potest* (es kann geschehen) vs. *factum est* (es ist geschehen)

Selektion als semantisches Indiz: Harris 1954; Kollokationen und *colligations* in der distributionellen Semantik: Firth 1957; Differenzierung fast synonyme Begriffe: Lehecka 2015; Verteilung und Funktion: Gries et al. 2009

Polysemie

- **Disambiguierung** durch
 - ▶ semantische **Relationen**: Macht im Spektrum zwischen Militär und Reichtum
 - ★ *auriferorum opes fluminum* (die Reichtümer goldener Flüsse)
 - ★ *Francorum opes penitus excisae* (die Truppen der Franken [wurden] vollkommen aufgerieben)

Disambiguierung durch semantische Relationen: Ayşe et al. 2011; Disambiguierung durch modifizierende Adjektive: Boleda et al. 2013; Kollokationsnetzwerke: Hamilton et al. 2016

Polysemie

● Disambiguierung durch

- ▶ semantische **Relationen**: Macht im Spektrum zwischen Militär und Reichtum
 - ★ *auriferorum opes fluminum* (die Reichtümer goldener Flüsse)
 - ★ **Francorum opes penitus excisae** (die Truppen der Franken [wurden] vollkommen aufgerieben)
- ▶ modifizierende **Adjektive**
 - ★ *Videor iam cernere non spoliis prouinciarum et extorto sociis auro, sed **hostilibus armis** captorumque regum catenis triumphum grauem* (Kriege mit Feinden)
 - ★ *quotiens equos, quotiens **emerita arma** mutasti!* (ausgediente Waffen)

Disambiguierung durch semantische Relationen: Ayşe et al. 2011; Disambiguierung durch modifizierende Adjektive: Boleda et al. 2013; Kollokationsnetzwerke: Hamilton et al. 2016

Polysemie

- **Disambiguierung** durch
 - ▶ semantische **Relationen**: Macht im Spektrum zwischen Militär und Reichtum
 - ★ *auriferorum opes fluminum* (die Reichtümer goldener Flüsse)
 - ★ *Francorum opes penitus excisae* (die Truppen der Franken [wurden] vollkommen aufgerieben)
 - ▶ modifizierende **Adjektive**
 - ★ *Videor iam cernere non spoliis prouinciarum et extorto sociis auro, sed hostilibus armis captorumque regum catenis triumphum grauem* (Kriege mit Feinden)
 - ★ *quotiens equos, quotiens emerita arma mutasti!* (ausgediente Waffen)
- **Quantifizierung** durch Kollokationsnetzwerke

Disambiguierung durch semantische Relationen: Ayşe et al. 2011; Disambiguierung durch modifizierende Adjektive: Boleda et al. 2013; Kollokationsnetzwerke: Hamilton et al. 2016



Distributionelle Semantische Modelle

Überblick

- sollen semantische Informationen aus Korpora extrahieren und **repräsentieren**

Semantische Relationen beim maschinellen Lernen: Carlson et al. 2010; Syntax vs. Semantik: Rich et al. 1991, Clark et al. 2019; Morphologie in DSMs: Gladkova et al. 2016; Vorteile von Korpora für semantischen Analysen: Gries et al. 2009

Überblick

- sollen semantische Informationen aus Korpora extrahieren und **repräsentieren**
- basierten früher auf Statistik, heute eher auf **maschinellem Lernen**

Semantische Relationen beim maschinellen Lernen: Carlson et al. 2010; Syntax vs. Semantik: Rich et al. 1991, Clark et al. 2019; Morphologie in DSMs: Gladkova et al. 2016; Vorteile von Korpora für semantischen Analysen: Gries et al. 2009

Überblick

- sollen semantische Informationen aus Korpora extrahieren und **repräsentieren**
- basierten früher auf Statistik, heute eher auf **maschinellern Lernen**
- nutzen mitunter ähnliche Informationen wie beim **menschlichen Spracherwerb**
 - ▶ Wissen über semantische Relationen
 - ▶ Interaktion von Syntax und Semantik

Semantische Relationen beim maschinellen Lernen: Carlson et al. 2010; Syntax vs. Semantik: Rich et al. 1991, Clark et al. 2019; Morphologie in DSMs: Gladkova et al. 2016; Vorteile von Korpora für semantischen Analysen: Gries et al. 2009

Überblick

- sollen semantische Informationen aus Korpora extrahieren und **repräsentieren**
- basierten früher auf Statistik, heute eher auf **maschinellern Lernen**
- nutzen mitunter ähnliche Informationen wie beim **menschlichen Spracherwerb**
 - ▶ Wissen über semantische Relationen
 - ▶ Interaktion von Syntax und Semantik
- zeigen neben semantischen oft auch **morphologische** Relationen, z.B. Demonstrativpronomina und ihre verschiedenen Flexionsformen
 - ▶ *illo* (jener, Abl.Sg.) → *illorum* (Gen.Pl.)
 - ▶ *illa* (Nom.Sg.) → *illius* (Gen.Sg.)

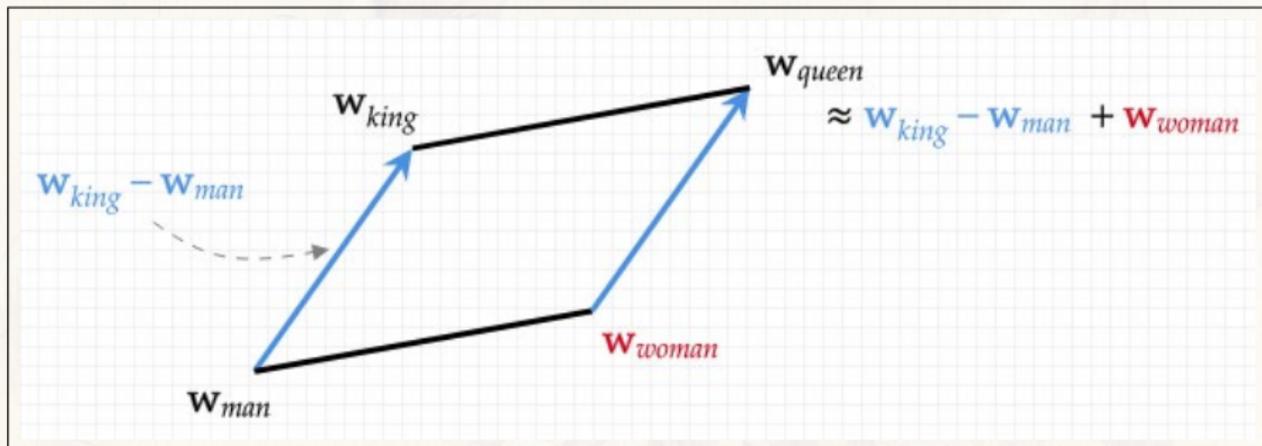
Semantische Relationen beim maschinellen Lernen: Carlson et al. 2010; Syntax vs. Semantik: Rich et al. 1991, Clark et al. 2019; Morphologie in DSMs: Gladkova et al. 2016; Vorteile von Korpora für semantischen Analysen: Gries et al. 2009

Schwächen

- übersehen offensichtliches **Allgemeinwissen**, das alle haben, aber über das niemand spricht
- reduzieren Polysemie auf eine **einzigste zentrale Bedeutung** pro Wort
- bilden seltene und **metaphorische** Wortbedeutungen schlecht ab
- unterscheiden nicht (fein genug) zwischen **Synonymie** und **syntagmatischer** Verwandtschaft
- ermöglichen nur bestimmte, **fallspezifische Analogien**

fehlendes Allgemeinwissen: Bruni et al. 2014; Nivellierung von Polysemie: Faruqui et al. 2016; fehlende Metaphorik: Grigonytė et al. 2010; Synonymie und Syntagmatik: Karan et al. 2012; eingeschränkte Analogien: Rogers et al. 2017

Wortanalogien



spadones - uiros + feminas = mulierculas

Entmannte - Männer + Frauen = (Frau im heiratsfähigen Alter)-chen



Maschinelles Lernen

Input: **In principio erat verbum et verbum erat apud deum et deus erat verbum.** (Am Anfang war das Wort...)

Aufgabe: Definiere die Bedeutung eines jeden Worts durch seinen linken und rechten **Nachbarn**.

Input: **In principio erat verbum et verbum erat apud deum et deus erat verbum.** (Am Anfang war das Wort...)

Aufgabe: Definiere die Bedeutung eines jeden Worts durch seinen linken und rechten **Nachbarn**.

- 1 Weise jedem Type einen **Identifikator** zu: In: 1, principio: 2, erat: 3, verbum: 4, et: 5, apud: 6, deum: 7, deus: 8

Input: **In principio erat verbum et verbum erat apud deum et deus erat verbum.** (Am Anfang war das Wort...)

Aufgabe: Definiere die Bedeutung eines jeden Worts durch seinen linken und rechten **Nachbarn**.

- 1 Weise jedem Type einen **Identifikator** zu: In: 1, principio: 2, erat: 3, verbum: 4, et: 5, apud: 6, deum: 7, deus: 8
- 2 Repräsentiere jedes Token als eine **Sequenz** von Identifikatoren: In = [1,2], principio = [1,2,3], erat₁ = [2,3,4], erat₂ = [4,3,6], erat₃ = [8,3,4]

Input: **In principio erat verbum et verbum erat apud deum et deus erat verbum.** (Am Anfang war das Wort...)

Aufgabe: Definiere die Bedeutung eines jeden Worts durch seinen linken und rechten **Nachbarn**.

- 1 Weise jedem Type einen **Identifikator** zu: In: 1, principio: 2, erat: 3, verbum: 4, et: 5, apud: 6, deum: 7, deus: 8
- 2 Repräsentiere jedes Token als eine **Sequenz** von Identifikatoren: In = [1,2], principio = [1,2,3], $erat_1 = [2,3,4]$, $erat_2 = [4,3,6]$, $erat_3 = [8,3,4]$
- 3 **Aggregiere** mehrfache Repräsentationen desselben Types:
 $erat = \frac{erat_1 + erat_2 + erat_3}{3}$

Input: **In principio erat verbum et verbum erat apud deum et deus erat verbum.** (Am Anfang war das Wort...)

Aufgabe: Definiere die Bedeutung eines jeden Worts durch seinen linken und rechten **Nachbarn**.

- 1 Weise jedem Type einen **Identifikator** zu: In: 1, principio: 2, erat: 3, verbum: 4, et: 5, apud: 6, deum: 7, deus: 8
- 2 Repräsentiere jedes Token als eine **Sequenz** von Identifikatoren: In = [1,2], principio = [1,2,3], $erat_1 = [2,3,4]$, $erat_2 = [4,3,6]$, $erat_3 = [8,3,4]$
- 3 **Aggregiere** mehrfache Repräsentationen desselben Types:
 $erat = \frac{erat_1 + erat_2 + erat_3}{3}$
- 4 Formuliere eine **Suchanfrage**: Was ist ein typischer Kontext von 'erat'? → verbum

Kleine Geschichte des Representation Learning

- Word2Vec (2013)
 - ▶ **1 Vektor pro Type**, feste Fenstergröße
 - ▶ gemeinsame Tokenisierung von Phrasen basierend auf **Assoziationsmaßen**

ELMo: Peters et al. 2018; Matrix-Faktorisierung: Pennington et al. 2014, Levy et al. 2014, Li et al. 2015; Word2Vec: Mikolov et al. 2013a, Mikolov et al. 2013b; BERT: Devlin et al. 2018; fastText: Bojanowski et al. 2017; FLAIR: Akbik et al. 2018; GPT-3: Brown et al. 2020

Kleine Geschichte des Representation Learning

- Word2Vec (2013)
 - ▶ **1 Vektor pro Type**, feste Fenstergröße
 - ▶ gemeinsame Tokenisierung von Phrasen basierend auf **Assoziationsmaßen**
- GloVe (2014): ergänzende **Matrix-Faktorisierung** der Kookkurrenzen

ELMo: Peters et al. 2018; Matrix-Faktorisierung: Pennington et al. 2014, Levy et al. 2014, Li et al. 2015; Word2Vec: Mikolov et al. 2013a, Mikolov et al. 2013b; BERT: Devlin et al. 2018; fastText: Bojanowski et al. 2017; FLAIR: Akbik et al. 2018; GPT-3: Brown et al. 2020

Kleine Geschichte des Representation Learning

- Word2Vec (2013)
 - ▶ **1 Vektor pro Type**, feste Fenstergröße
 - ▶ gemeinsame Tokenisierung von Phrasen basierend auf **Assoziationsmaßen**
- GloVe (2014): ergänzende **Matrix-Faktorisierung** der Kookkurrenzen
- fastText (2016): Berücksichtigung von **Morphemen**

ELMo: Peters et al. 2018; Matrix-Faktorisierung: Pennington et al. 2014, Levy et al. 2014, Li et al. 2015; Word2Vec: Mikolov et al. 2013a, Mikolov et al. 2013b; BERT: Devlin et al. 2018; fastText: Bojanowski et al. 2017; FLAIR: Akbik et al. 2018; GPT-3: Brown et al. 2020

Kleine Geschichte des Representation Learning

- Word2Vec (2013)
 - ▶ **1 Vektor pro Type**, feste Fenstergröße
 - ▶ gemeinsame Tokenisierung von Phrasen basierend auf **Assoziationsmaßen**
- GloVe (2014): ergänzende **Matrix-Faktorisierung** der Kookkurrenzen
- fastText (2016): Berücksichtigung von **Morphemen**
- FLAIR (2018): Embeddings auf **Zeichenebene**

ELMo: Peters et al. 2018; Matrix-Faktorisierung: Pennington et al. 2014, Levy et al. 2014, Li et al. 2015; Word2Vec: Mikolov et al. 2013a, Mikolov et al. 2013b; BERT: Devlin et al. 2018; fastText: Bojanowski et al. 2017; FLAIR: Akbik et al. 2018; GPT-3: Brown et al. 2020

Kleine Geschichte des Representation Learning

- Word2Vec (2013)
 - ▶ **1 Vektor pro Type**, feste Fenstergröße
 - ▶ gemeinsame Tokenisierung von Phrasen basierend auf **Assoziationsmaßen**
- GloVe (2014): ergänzende **Matrix-Faktorisierung** der Kookkurrenzen
- fastText (2016): Berücksichtigung von **Morphemen**
- FLAIR (2018): Embeddings auf **Zeichenebene**
- ELMo (2018): Modellierung vorwärts und **rückwärts**

ELMo: Peters et al. 2018; Matrix-Faktorisierung: Pennington et al. 2014, Levy et al. 2014, Li et al. 2015; Word2Vec: Mikolov et al. 2013a, Mikolov et al. 2013b; BERT: Devlin et al. 2018; fastText: Bojanowski et al. 2017; FLAIR: Akbik et al. 2018; GPT-3: Brown et al. 2020

Kleine Geschichte des Representation Learning

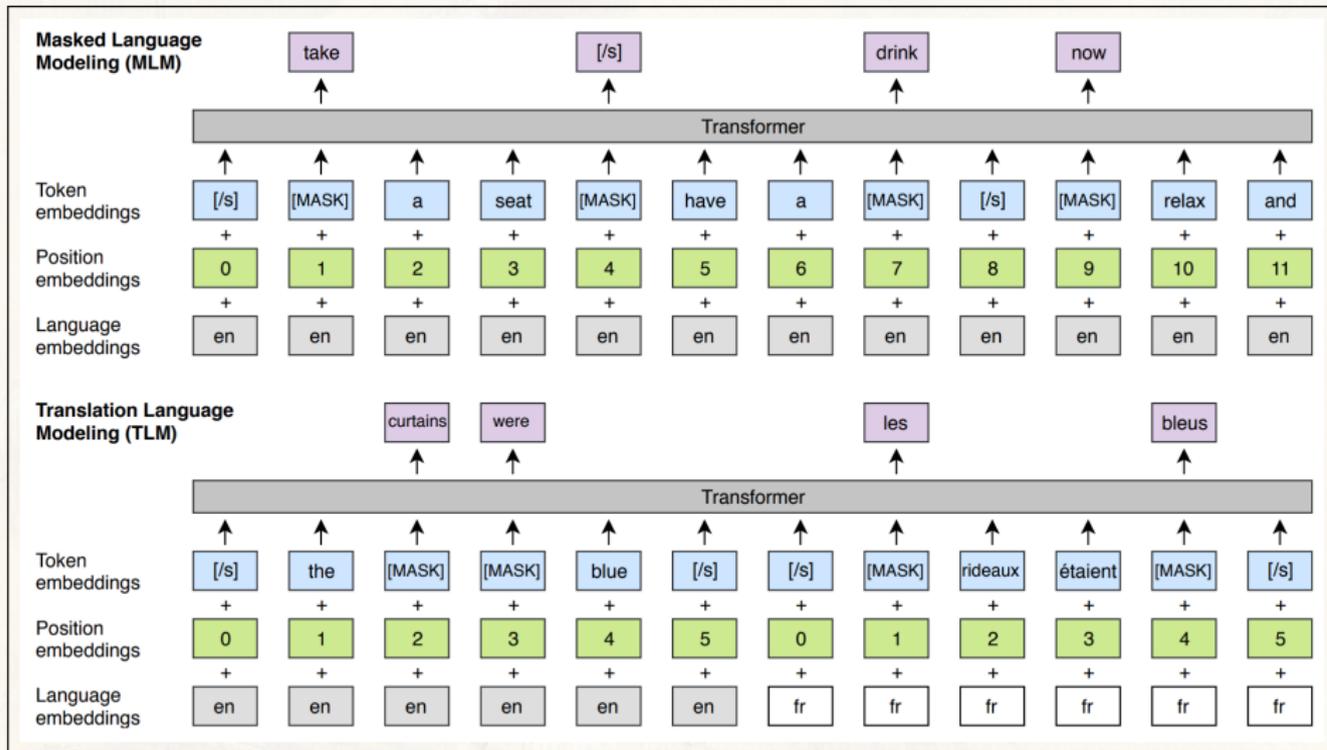
- Word2Vec (2013)
 - ▶ **1 Vektor pro Type**, feste Fenstergröße
 - ▶ gemeinsame Tokenisierung von Phrasen basierend auf **Assoziationsmaßen**
- GloVe (2014): ergänzende **Matrix-Faktorisierung** der Kookkurrenzen
- fastText (2016): Berücksichtigung von **Morphemen**
- FLAIR (2018): Embeddings auf **Zeichenebene**
- ELMo (2018): Modellierung vorwärts und **rückwärts**
- BERT (2018): **attention**-basierte Berücksichtigung von Kontext

ELMo: Peters et al. 2018; Matrix-Faktorisierung: Pennington et al. 2014, Levy et al. 2014, Li et al. 2015; Word2Vec: Mikolov et al. 2013a, Mikolov et al. 2013b; BERT: Devlin et al. 2018; fastText: Bojanowski et al. 2017; FLAIR: Akbik et al. 2018; GPT-3: Brown et al. 2020

Kleine Geschichte des Representation Learning

- Word2Vec (2013)
 - ▶ **1 Vektor pro Type**, feste Fenstergröße
 - ▶ gemeinsame Tokenisierung von Phrasen basierend auf **Assoziationsmaßen**
- GloVe (2014): ergänzende **Matrix-Faktorisierung** der Kookkurrenzen
- fastText (2016): Berücksichtigung von **Morphemen**
- FLAIR (2018): Embeddings auf **Zeichenebene**
- ELMo (2018): Modellierung vorwärts und **rückwärts**
- BERT (2018): **attention**-basierte Berücksichtigung von Kontext
- GPT-3 (2020): $10 + 20 = 30$ 🧐

ELMo: Peters et al. 2018; Matrix-Faktorisierung: Pennington et al. 2014, Levy et al. 2014, Li et al. 2015; Word2Vec: Mikolov et al. 2013a, Mikolov et al. 2013b; BERT: Devlin et al. 2018; fastText: Bojanowski et al. 2017; FLAIR: Akbik et al. 2018; GPT-3: Brown et al. 2020

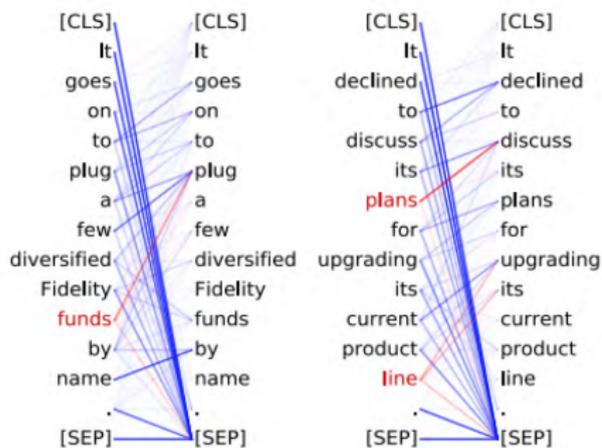


Quelle: Lample et al. 2019; kontextualisierte Embeddings: Wiedemann et al. 2019

Attention

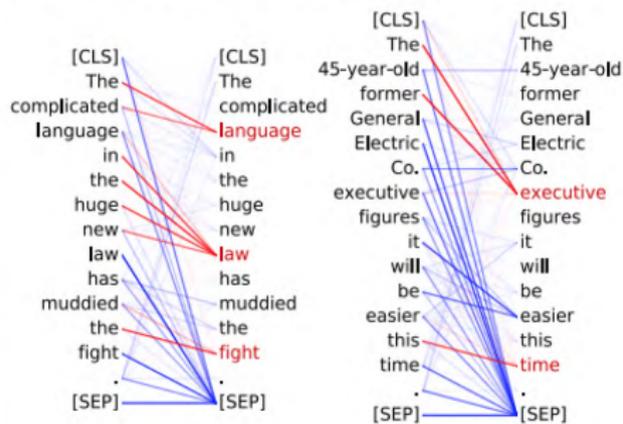
Head 8-10

- **Direct objects** attend to their verbs
- 86.8% accuracy at the dobj relation



Head 8-11

- **Noun modifiers** (e.g., determiners) attend to their noun
- 94.3% accuracy at the det relation



Quelle: Clark et al. 2019

End-to-end learning

- stages of NLP (Dale 2010, p.4)

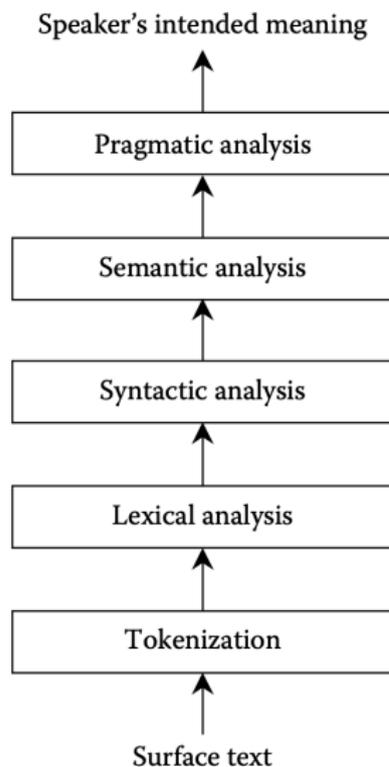


FIGURE 1.1 The stages of analysis in processing natural language.

End-to-end learning

- stages of NLP (Dale 2010, p.4)

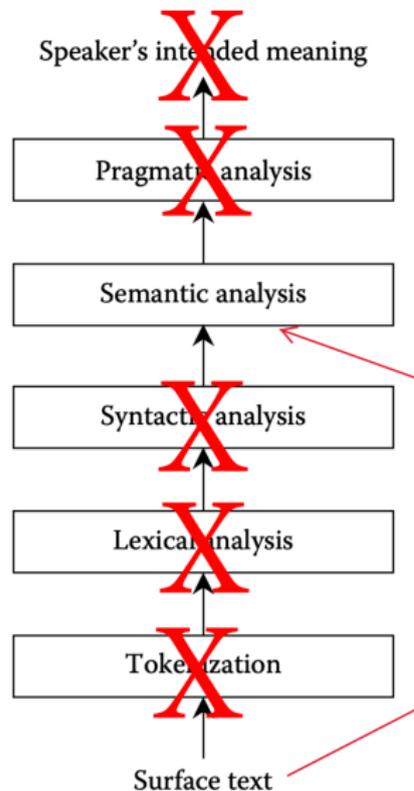


FIGURE 1.1 The stages of analysis in processing natural language.

Stufen der semantischen Analyse und ihre Berücksichtigung beim Maschinellen Lernen

- Konstruktion: Analyse von **Lexik und Syntax**
 - ▶ Treebanks
 - ▶ Assoziationsmaße / Kollokationen

Analysestufen: Pinkal 1993; Treebanks: Haug et al. 2008, Bamman et al. 2011, Cecchini et al. 2018; Assoziationsmaße: Gries 2013, Gries et al. 2019; Word Sense Disambiguation: Bartunov et al. 2016; Holzwegeffekt: Farmer et al. 2011; Question Natural Language Inference: Jiao et al. 2019; ergänzende Materialien und Regeln zur Inferenz: Punyakanok et al. 2008, Ono et al. 2015

Stufen der semantischen Analyse und ihre Berücksichtigung beim Maschinellen Lernen

- Konstruktion: Analyse von **Lexik und Syntax**
 - ▶ Treebanks
 - ▶ Assoziationsmaße / Kollokationen
- Resolution: **Disambiguierung**
 - ▶ Word Sense Disambiguation
 - ▶ Holzwegeffekt

Analysestufen: Pinkal 1993; Treebanks: Haug et al. 2008, Bamman et al. 2011, Cecchini et al. 2018; Assoziationsmaße: Gries 2013, Gries et al. 2019; Word Sense Disambiguation: Bartunov et al. 2016; Holzwegeffekt: Farmer et al. 2011; Question Natural Language Inference: Jiao et al. 2019; ergänzende Materialien und Regeln zur Inferenz: Punyakanok et al. 2008, Ono et al. 2015

Stufen der semantischen Analyse und ihre Berücksichtigung beim Maschinellen Lernen

- Konstruktion: Analyse von **Lexik und Syntax**
 - ▶ Treebanks
 - ▶ Assoziationsmaße / Kollokationen
- Resolution: **Disambiguierung**
 - ▶ Word Sense Disambiguation
 - ▶ Holzwegeffekt
- Auswertung: Deduktion und **Inferenz** mittels Weltwissen
 - ▶ ergänzende Materialien, z.B. Thesauri
 - ▶ Question Natural Language Inference: Liefert ein gegebenener **Kontext** die Antwort zu einer gegebenen **Frage**?

Analysestufen: Pinkal 1993; Treebanks: Haug et al. 2008, Bamman et al. 2011, Cecchini et al. 2018; Assoziationsmaße: Gries 2013, Gries et al. 2019; Word Sense Disambiguation: Bartunov et al. 2016; Holzwegeffekt: Farmer et al. 2011; Question Natural Language Inference: Jiao et al. 2019; ergänzende Materialien und Regeln zur Inferenz: Punyakanok et al. 2008, Ono et al. 2015

The first recorded travels by Europeans to China and back date from this time. The most famous traveler of the period was the Venetian Marco Polo, whose account of his trip to "Cambaluc," the capital of the Great Khan, and of life there astounded the people of Europe. The account of his travels, *Il milione* (or, *The Million*, known in English as the *Travels of Marco Polo*), appeared about the year 1299. Some argue over the accuracy of Marco Polo's accounts due to the lack of mentioning the Great Wall of China, tea houses, which would have been a prominent sight since Europeans had yet to adopt a tea culture, as well the practice of foot binding by the women in capital of the Great Khan. Some suggest that Marco Polo acquired much of his knowledge through contact with Persian traders since many of the places he named were in Persian.

How did some suspect that Polo learned about China instead of by actually visiting it?

Answer: through contact with Persian traders

Word Sense Disambiguation

| Context: He plays a pianist in the film | | | | | |
|--|--------|--------|--------|--------|--------|
| Glosses | Pass 1 | Pass 2 | Pass 3 | Pass 4 | Pass 5 |
| g_1 : participate in games or sport | | | | | |
| g_2 : perform music on a instrument | | | | | |
| g_3 : act a role or part | | | | | |

Table 2: An example of attention weights in the memory module within 5 passes. Darker colors mean that the attention weight is higher. Case studies show that the proposed multi-pass operation can recognize the correct sense by enlarging the attention gap between correct senses and incorrect ones.

Quelle: Luo et al. 2018

Grundsätze der Optimierung

- stark abhängig von
 - ▶ **Sprache**
 - ▶ geplante/notwendige **Verarbeitungsschritte**
 - ▶ Umfang von
 - ★ **Trainingsdaten**
 - ★ **Modell**
 - ★ vorhandene **Hardware**

unterschiedliche Konfiguration je nach Bedingungen: Dobó 2019; Umfang der Trainingsdaten: McGillivray 2013, Herbelot et al. 2015, Hestness et al. 2017, Karakanta et al. 2018; Umfang der Modelle vs. Trainingsdaten: Kaplan et al. 2020

Grundsätze der Optimierung

- stark abhängig von
 - ▶ **Sprache**
 - ▶ geplante/notwendige **Verarbeitungsschritte**
 - ▶ Umfang von
 - ★ **Trainingsdaten**
 - ★ **Modell**
 - ★ vorhandene **Hardware**
- benötigt gute Evaluation, d.h. sinnvoll strukturierte Testdatensätze
 - ▶ Berücksichtigung von **Bias**, ggf. entsprechende Normalisierung
 - ▶ Abdeckung verschiedener **Zielszenarien** in einem einheitlichen Framework

unterschiedliche Konfiguration je nach Bedingungen: Dobó 2019; Umfang der Trainingsdaten: McGillivray 2013, Herbelot et al. 2015, Hestness et al. 2017, Karakanta et al. 2018; Umfang der Modelle vs. Trainingsdaten: Kaplan et al. 2020

Konkrete Optimierungsmaßnahmen

- **byte pair encoding** zur flexiblen Tokenisierung
- **memorization filtering** gegen Auswendiglernen von bestimmten Mustern
- **domain adaptation** mittels abstrakter Zwischenrepräsentationen
- **curriculum learning** für schrittweise härtere Herausforderungen
- **knowledge distillation** zur Verkleinerung existierender Modelle
- **parameter sharing** zur effizienten Kombination mehrerer Netzwerke
- **teacher forcing** zur Korrektur von zwischenzeitigem Output

memorization filtering: Gupta et al. 2018; domain adaptation: Glorot et al. 2011;
curriculum learning: Graves et al. 2017; knowledge distillation: Jiao et al. 2019;
parameter sharing: Dehghani et al. 2019, Lan et al. 2020; byte pair encoding: Sennrich et al. 2016; teacher forcing: Su et al. 2018



Geplante Anwendung

Begriffe zu einem Thema

- Antike: weniger **Trainingsdaten** verfügbar als für moderne Sprachen
→ Fokus auf
 - ▶ hohe Rechenleistung (**HU-Hardware**)
 - ▶ Modelle mit **vielen Parametern**

Begriffe zu einem Thema

- Antike: weniger **Trainingsdaten** verfügbar als für moderne Sprachen
→ Fokus auf
 - ▶ hohe Rechenleistung (**HU-Hardware**)
 - ▶ Modelle mit **vielen Parametern**
- ① antikes **Ausgangswort** identifizieren
 - ▶ direkt als Input
 - ▶ per Übersetzung mittels **Wörterbuchanbindung**

Begriffe zu einem Thema

- Antike: weniger **Trainingsdaten** verfügbar als für moderne Sprachen
→ Fokus auf
 - ▶ hohe Rechenleistung (**HU-Hardware**)
 - ▶ Modelle mit **vielen Parametern**
- ① antikes **Ausgangswort** identifizieren
 - ▶ direkt als Input
 - ▶ per Übersetzung mittels **Wörterbuchanbindung**
- ② **Varianten** des Worts finden
 - ▶ Lemmatisierung
 - ▶ morphologische Analyse
 - ▶ Synonymik

Begriffe zu einem Thema

- Antike: weniger **Trainingsdaten** verfügbar als für moderne Sprachen
→ Fokus auf
 - ▶ hohe Rechenleistung (**HU-Hardware**)
 - ▶ Modelle mit **vielen Parametern**
- ① antikes **Ausgangswort** identifizieren
 - ▶ direkt als Input
 - ▶ per Übersetzung mittels **Wörterbuchanbindung**
- ② **Varianten** des Worts finden
 - ▶ Lemmatisierung
 - ▶ morphologische Analyse
 - ▶ Synonymik
- ③ höchste **Kosinus-Ähnlichkeit** für Vektoren aller Ausgangswörter

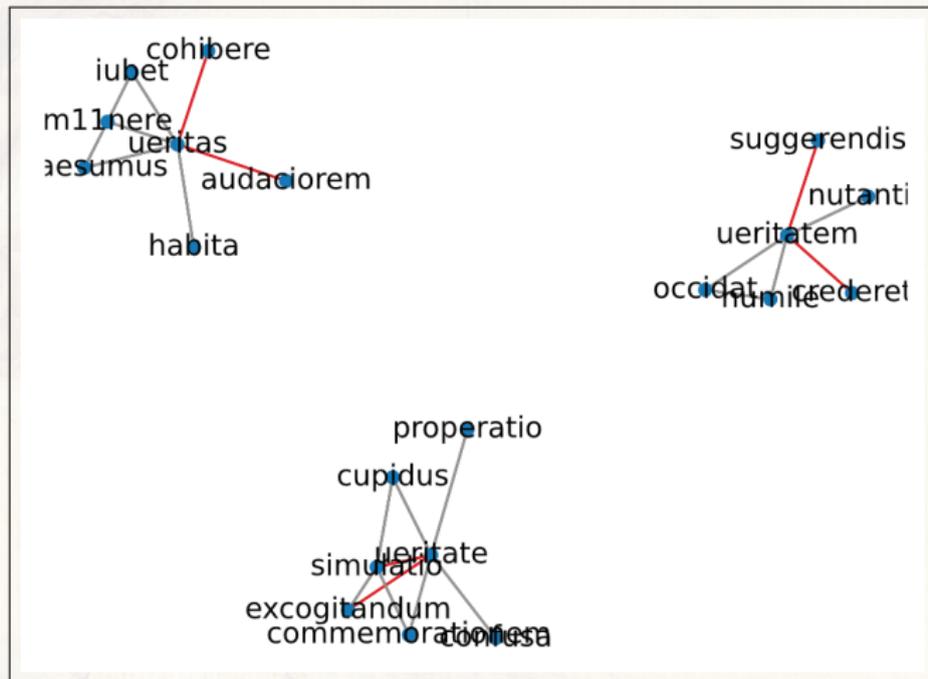
Begriffe zu einem Thema

- Antike: weniger **Trainingsdaten** verfügbar als für moderne Sprachen
→ Fokus auf
 - ▶ hohe Rechenleistung (**HU-Hardware**)
 - ▶ Modelle mit **vielen Parametern**
- ① antikes **Ausgangswort** identifizieren
 - ▶ direkt als Input
 - ▶ per Übersetzung mittels **Wörterbuchanbindung**
- ② **Varianten** des Worts finden
 - ▶ Lemmatisierung
 - ▶ morphologische Analyse
 - ▶ Synonymik
- ③ höchste **Kosinus-Ähnlichkeit** für Vektoren aller Ausgangswörter
- ④ **Querverbindungen** zwischen Zielwörtern ermitteln

Begriffe zu einem Thema

- Antike: weniger **Trainingsdaten** verfügbar als für moderne Sprachen
→ Fokus auf
 - ▶ hohe Rechenleistung (**HU-Hardware**)
 - ▶ Modelle mit **vielen Parametern**
- ① antikes **Ausgangswort** identifizieren
 - ▶ direkt als Input
 - ▶ per Übersetzung mittels **Wörterbuchanbindung**
- ② **Varianten** des Worts finden
 - ▶ Lemmatisierung
 - ▶ morphologische Analyse
 - ▶ Synonymik
- ③ höchste **Kosinus-Ähnlichkeit** für Vektoren aller Ausgangswörter
- ④ **Querverbindungen** zwischen Zielwörtern ermitteln
- ⑤ Visualisierung als interaktives **Netzwerk**

- simulatio
(*Vorwand*)
- crederet
(*glauben*)
- suggerendis
(*hinzufügen, zu etw. raten*)
- excogitandum
(*sich etw. ausdenken*)
- cohibere
(*zurückhalten*)
- audaciorem
(*tollkühn*)



Intertextuelle Parallelen

- 1 antike **Ausgangswörter** identifizieren
 - ▶ Lemmatisierung + morphologische Analyse
 - ▶ Regular Expression

Intertextuelle Parallelen

- 1 antike **Ausgangswörter** identifizieren
 - ▶ Lemmatisierung + morphologische Analyse
 - ▶ Regular Expression
- 2 **Umwandlung** in Vektoren

Intertextuelle Parallelen

- 1 antike **Ausgangswörter** identifizieren
 - ▶ Lemmatisierung + morphologische Analyse
 - ▶ Regular Expression
- 2 **Umwandlung** in Vektoren
- 3 **Aggregation** zu einzeltem Zielvektor

Intertextuelle Parallelen

- 1 antike **Ausgangswörter** identifizieren
 - ▶ Lemmatisierung + morphologische Analyse
 - ▶ Regular Expression
- 2 **Umwandlung** in Vektoren
- 3 **Aggregierung** zu einzeltem Zielvektor
- 4 Schritte 1-3 wiederholen für **jeden Satz** in einem Korpus

Intertextuelle Parallelen

- 1 antike **Ausgangswörter** identifizieren
 - ▶ Lemmatisierung + morphologische Analyse
 - ▶ Regular Expression
- 2 **Umwandlung** in Vektoren
- 3 **Aggregation** zu einzeltem Zielvektor
- 4 Schritte 1-3 wiederholen für **jeden Satz** in einem Korpus
- 5 Berechnung der **Kosinus-Ähnlichkeit** zwischen Satzvektoren und Anfragevektor

Intertextuelle Parallelen

- 1 antike **Ausgangswörter** identifizieren
 - ▶ Lemmatisierung + morphologische Analyse
 - ▶ Regular Expression
- 2 **Umwandlung** in Vektoren
- 3 **Aggregation** zu einzeltem Zielvektor
- 4 Schritte 1-3 wiederholen für **jeden Satz** in einem Korpus
- 5 Berechnung der **Kosinus-Ähnlichkeit** zwischen Satzvektoren und Anfragevektor
- 6 Optional: Kombination mit anderen Faktoren, z.B. **lexikalische Überschneidung**

Intertextuelle Parallelen

- 1 antike **Ausgangswörter** identifizieren
 - ▶ Lemmatisierung + morphologische Analyse
 - ▶ Regular Expression
- 2 **Umwandlung** in Vektoren
- 3 **Aggregierung** zu einzeltem Zielvektor
- 4 Schritte 1-3 wiederholen für **jeden Satz** in einem Korpus
- 5 Berechnung der **Kosinus-Ähnlichkeit** zwischen Satzvektoren und Anfragevektor
- 6 Optional: Kombination mit anderen Faktoren, z.B. **lexikalische Überschneidung**
- 7 **Ranking**, Ausgabe als Trefferliste

Inhaltlich ähnliche Textstellen

Input: *uera* (Wahres)

haec pro imperatore nostro in quae 5 sint uerba suscepta ope¹-ae pretium est adnotare : rsi bene rem publicam et ex utilitate omnium * * * digna uota quae semper suscipiantur semper -que soluuntur .

ut haec esse uera credamus quae mendacilis uatum in plausus aptata cauearum .

merito necne , neutram in partem decernere audeamus , quia nec fas est affirmationi tuae derogare et onerosum confiteri uera esse quae de nobis , praesertim tam magnifica , dixisti .

scis enim uhi uera principis , ubi sempiterna sit gloria .

quae perpetuae et constantis benit¹olentiae prima **fiducia** est , **certissima** ui¹-tut¹im et max.ima principis , 11eritas , numquam in animo esse **susp¹icacem** te audiui .

dico enim magna **certa** -que **fiducia** : digna est hac quoque glo¹-ia uestra pietas , et potest eam praestare felicitas .

o tego properationem militum iuui , ego fugam hostium praepediui , ego Maxim¹im in muros coegi et quem uos **cogebatis** mori uiuum domino reseruau¹ : non uideo quod possit esse iudicium quin , cum ista pars imputet tyranni fugam , illa custodiam , res publica

20 legii religionem uererer , si is , qui semper 1-es a numine tuo gestas praedicare solitus essen¹ , haec tanto maiora pristinis silentio **praeterirem** , per quae non pars aliqua seruata , sed uniuersa sibi est res publica

neque enim parui negotii est imperatorem totius or bis pro se peculiariter rogare , sub tantae maiestatis aspectu perficere frontem , uultum componere , 20 confirmare animum , uerba concipi pere , **intrepidante**¹ : dicere , apte desinere , expectare responsum .

quibus enim m¹1neribus aequari 20 haec cura principis , haec patientia potest , qua meruisti ut ille sibi nimium fortis ac p¹-ope durus uideretur ?

est enim haec demum uera felicitas ao

haec est uera pietas , haec uoluptas conseruandi generis humani , exemplum dare gentibus ad matrimonia studiosius expetenda et liberos educandos , ut substituendis so successioneibus singulorum nihil obsit unumquemque esse mortalem , cum immortalis sit omnium posteritate res publica .

etenim ceterae uirtutes et bona cetera processu aetatis eueniunt , fortitudo a.nnis accedentibus roboratur , continentia disciplinae praeceptis traditur , iustitia **cognitio** -ne iuris addiscitur , 6 ipsa denique illa quae uidetur rerum omnium domina esse **sapientia** perspectis homin¹m moribus et exploratis rerum docetur euentis : solae cum nascentibus pariter oriuntur pietas atque felicitas ; naturalia sunt enim animorum bona et praemia fatorum .

ibi adulescentes optimi discant , nobis quasi sollemne carmen praefantibus , maximorum principum¹ **facta** celebrare (quis enim melior usus est eloquentiae ?

neque enim **fabula** est de **licentia** poetarum nec **opinio** de fama ueterum saeculorum , sed **manifesta** res et **probata** , 20 sicut hodie -que testatur Herculis ara maxima et Herculei sacri custos familia Pinaria , principem illum tui generis ac nominis Pallantea moenia adisse ui

magnus uterque honor uester , sit¹e eosdem maxime diligit quos scit uobis esse carissimos siue illis neminem praefert , quamuis aliquem magis amet : proposita sunt senioribus p¹-aemia , iuuenib¹is exempla : adeant , 20 frequentent securas tandem ac patentes domos : quisquis **probatus** senatui uiros suscipit , hic maxime principem promeretur .

etenim ut nomj -ne Augusti admonemur eius cui primum dicatum est , ita haec optimi appellatio numquam memoriae l¹ominum sine te recurret , quotiensque poster¹ nostri optimum aliquem uocare **cogentur** , so totiens recordabuntur quis me¹-uerit uocari .

Genrespezifischer Sprachgebrauch

- 1 verschiedene **Analysedimensionen** festlegen
 - ▶ **Dependenz**grammatik
 - ▶ Assoziationsmaße / **Kookkurrenzen**
 - ▶ Paragraph- bzw. **Dokumentvektoren**

Paragraphvektoren: Le et al. 2014, Dai et al. 2015

Genrespezifischer Sprachgebrauch

- 1 verschiedene **Analysedimensionen** festlegen
 - ▶ **Dependenz**grammatik
 - ▶ Assoziationsmaße / **Kookkurrenzen**
 - ▶ Paragraph- bzw. **Dokumentvektoren**
- 2 Basis- und Referenzkorpora mittels **Metadaten** auswählen
 - ▶ Autorschaft
 - ▶ Entstehungszeit
 - ▶ Genre

Paragraphvektoren: Le et al. 2014, Dai et al. 2015

Genrespezifischer Sprachgebrauch

- 1 verschiedene **Analysedimensionen** festlegen
 - ▶ **Dependenz**grammatik
 - ▶ Assoziationsmaße / **Kookkurrenzen**
 - ▶ Paragraph- bzw. **Dokumentvektoren**
- 2 Basis- und Referenzkorpora mittels **Metadaten** auswählen
 - ▶ Autorschaft
 - ▶ Entstehungszeit
 - ▶ Genre
- 3 **Verarbeitung**
 - ▶ Dependency Parsing / Klassifikation
 - ▶ Statistische Algorithmen
 - ▶ Maschinelles Lernen

Paragraphvektoren: Le et al. 2014, Dai et al. 2015

Genrespezifischer Sprachgebrauch

- 1 verschiedene **Analysedimensionen** festlegen
 - ▶ **Dependenz**grammatik
 - ▶ Assoziationsmaße / **Kookkurrenzen**
 - ▶ Paragraph- bzw. **Dokumentvektoren**
- 2 Basis- und Referenzkorpora mittels **Metadaten** auswählen
 - ▶ Autorschaft
 - ▶ Entstehungszeit
 - ▶ Genre
- 3 **Verarbeitung**
 - ▶ Dependency Parsing / Klassifikation
 - ▶ Statistische Algorithmen
 - ▶ Maschinelles Lernen
- 4 **Aufbereitung**
 - ▶ Diagramme
 - ▶ Tabellen
 - ▶ Querverweise

Paragraphvektoren: Le et al. 2014, Dai et al. 2015

Mögliche Probleme

- **Urheberrecht**

- ▶ Primärliteratur
- ▶ Sekundärliteratur
- ▶ Datensätze

Generell zu FAIRen Daten: Wilkinson et al. 2016

Mögliche Probleme

- **Urheberrecht**

- ▶ Primärliteratur
- ▶ Sekundärliteratur
- ▶ Datensätze

- digitale **Verfügbarkeit**

- ▶ Originaltexte
- ▶ Wörterbücher
- ▶ Übersetzungen

Generell zu FAIRen Daten: Wilkinson et al. 2016

Mögliche Probleme

- **Urheberrecht**

- ▶ Primärliteratur
- ▶ Sekundärliteratur
- ▶ Datensätze

- digitale **Verfügbarkeit**

- ▶ Originaltexte
- ▶ Wörterbücher
- ▶ Übersetzungen

- **Modellierung** komplexer Phänomene wie Intertextualität

Generell zu FAIRen Daten: Wilkinson et al. 2016

Mögliche Probleme

- **Urheberrecht**

- ▶ Primärliteratur
- ▶ Sekundärliteratur
- ▶ Datensätze

- digitale **Verfügbarkeit**

- ▶ Originaltexte
- ▶ Wörterbücher
- ▶ Übersetzungen

- **Modellierung** komplexer Phänomene wie Intertextualität

- **Evaluation**

- ▶ kaum vorhandene Testdatensätze
- ▶ mögliche **Alternativen**
 - ★ Digitalisierung älterer Ressourcen (z.B. Lexika)
 - ★ eigene Kuratierung durch manuelle Annotation

Generell zu FAIRen Daten: Wilkinson et al. 2016

Zu entwickelnde Testdatensätze

● Intertextualität

- ▶ Ausgangspunkt: Sequenz antiker Wörter, ggf. gewichtet
- ▶ Referenzpunkt: andere, potentiell relevante Sequenz antiker Wörter
- ▶ Entscheidung: graduell zwischen 0 und 1 (Relevanz)
- ▶ offene Fragen
 - ★ **Modellierung** auf mehreren Ebenen
 - ★ **Interaktion** zwischen benachbarten potentiellen Referenzpunkten

Evaluation distributioneller Semantik: Baroni et al. 2011

Zu entwickelnde Testdatensätze

● Intertextualität

- ▶ Ausgangspunkt: Sequenz antiker Wörter, ggf. gewichtet
- ▶ Referenzpunkt: andere, potentiell relevante Sequenz antiker Wörter
- ▶ Entscheidung: graduell zwischen 0 und 1 (Relevanz)
- ▶ offene Fragen
 - ★ **Modellierung** auf mehreren Ebenen
 - ★ **Interaktion** zwischen benachbarten potentiellen Referenzpunkten

● Wortfelder (Synonyme, Kollokationen...)

- ▶ Ausgangspunkt: einzelnes antikes Wort mit disambigierter Bedeutung
- ▶ Referenzpunkt: andere, potentiell relevante antike Wörter mit disambigierter Bedeutung
- ▶ Entscheidung: graduell zwischen 0 und 1 (Relevanz)
- ▶ offene Fragen
 - ★ **Referenzkorpus** als Einstiegspunkt / erste Orientierung
 - ★ **Balancing** nach Frequenz, Ambiguität, etc.

*Adhuc praesentia tua fruimur,
et iam reditum desideramus.*

Noch genießen wir deine Anwesenheit,
und schon wünschen wir uns deine baldige Rückkehr.

Item eiusdem magistri Mamertini genethiacus Maximiani Augusti – Rede zur Feier der Geburtstage der beiden Kaiser Diokletian und Maximian aus dem Jahr 290/291

Literatur I

- Akbik, A., D. Blythe, and R. Vollgraf (2018). “Contextual String Embeddings for Sequence Labeling”. In: *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 1638–1649.
- Ayşe, Ş., O. Zeynep, and P. İlknur (June 2011). “Extraction of Semantic Word Relations in Turkish from Dictionary Definitions”. In: *Proceedings of the ACL 2011 Workshop on Relational Models of Semantics*. Portland, Oregon, USA: Association for Computational Linguistics, pp. 11–18.
- Bamman, D. and G. Crane (2011). “The Ancient Greek and Latin Dependency Treebanks [AGLDT]”. In: *Language Technology for Cultural Heritage*. Springer, pp. 79–98.
- Baroni, M. and A. Lenci (2011). “How We BLESSed Distributional Semantic Evaluation”. In: *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*. Association for Computational Linguistics, pp. 1–10.

Literatur II

- Bartunov, S., D. Kondrashkin, A. Osokin, and D. Vetrov (2016). “Breaking Sticks and Ambiguities with Adaptive Skip-Gram”. In: *Artificial Intelligence and Statistics*, pp. 130–138.
- Bojanowski, P., E. Grave, A. Joulin, and T. Mikolov (June 2017). “Enriching Word Vectors with Subword Information [fastText]”. In: *arXiv:1607.04606 [cs]*. arXiv: 1607.04606 [cs].
- Boleda, G., M. Baroni, and L. McNally (2013). “Intensionality Was Only Alleged: On Adjective–Noun Composition in Distributional Semantics”. In: *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013): Long Papers; 2013 Mar 20-22; Postdam, Germany. Stroudsburg (USA): Association for Computational Linguistics (ACL)*. ACL (Association for Computational Linguistics), pp. 35–46.
- Brown, T. B. et al. (June 2020). “Language Models Are Few-Shot Learners”. In: *arXiv:2005.14165 [cs]*. arXiv: 2005.14165 [cs].

Literatur III

- Bruni, E., N.-K. Tran, and M. Baroni (2014). “Multimodal Distributional Semantics”. In: *J. Artif. Intell. Res. (JAIR)* 49.2014, pp. 1–47.
- Carlson, A. et al. (2010). “Toward an Architecture for Never-Ending Language Learning”. In: *Twenty-Fourth AAAI Conference on Artificial Intelligence*, pp. 1306–1313.
- Cecchini, F. M., M. Passarotti, P. Marongiu, and D. Zeman (2018). “Challenges in Converting the Index Thomisticus Treebank into Universal Dependencies [ITTB]”. In: *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pp. 27–36.
- Clark, K., U. Khandelwal, O. Levy, and C. D. Manning (June 2019). “What Does BERT Look At? An Analysis of BERT’s Attention”. In: *arXiv:1906.04341 [cs]*. arXiv: 1906.04341 [cs].

Literatur IV

- Cordes, L. (2020). “Wenn Fiktionen Fakten schaffen. Faktuales und fiktionales Erzählen in den spätantiken Panegyrici Latini”. *Deutsch. In: Faktuales und fiktionales Erzählen II. Geschichte - Medien - Praktiken.* Ed. by D. Breitenwischer, H.-M. Häger, and J. Menninger. Ergon-Verlag, pp. 31–56. DOI: 10.5771/9783956505126-31.
- Dai, A. M., C. Olah, and Q. V. Le (July 2015). “Document Embedding with Paragraph Vectors”. In: *arXiv:1507.07998 [cs]*. arXiv: 1507.07998 [cs].
- Dale, R. (2010). “Classical Approaches to Natural Language Processing”. In: *Handbook of Natural Language Processing.* Ed. by N. Indurkha and F. J. Damerau. Second. CRC Press, Taylor & Francis Group, pp. 3–7.
- Dehghani, M. et al. (Mar. 2019). “Universal Transformers”. In: *arXiv:1807.03819 [cs, stat]*. arXiv: 1807.03819 [cs, stat].

Literatur V

- Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2018). “Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding”. In: *arXiv preprint arXiv:1810.04805*.
- Dobó, A. (2019). “A Comprehensive Analysis of the Parameters in the Creation and Comparison of Feature Vectors in Distributional Semantic Models for Multiple Languages”. *PhD Thesis. szte*.
- Doran, D., S. Schulz, and T. R. Besold (Oct. 2017). “What Does Explainable AI Really Mean? A New Conceptualization of Perspectives”. en. In: *arXiv:1710.00794 [cs]*, p. 8. *arXiv: 1710.00794 [cs]*.
- Farmer, T., A. Fine, and F. Jaeger (2011). “Implicit Context-Specific Learning Leads to Rapid Shifts in Syntactic Expectations”. In: *Proceedings of the Cognitive Science Society*. Vol. 33, pp. 2055–2060.
- Faruqui, M., Y. Tsvetkov, P. Rastogi, and C. Dyer (2016). “Problems with Evaluation of Word Embeddings Using Word Similarity Tasks”. In: *arXiv preprint arXiv:1605.02276*.

Literatur VI

- Firth, J. R. (1957). “A Synopsis of Linguistic Theory 1930-55”. In: *Studies in Linguistic Analysis*. Ed. by J. R. Firth. Oxford: Blackwell, pp. 1–32.
- Frolov, S. and A. Wright (2011). “Homeric and Ancient Near Eastern Intertextuality in 1 Samuel 17”. In: *Journal of Biblical Literature* 130.3, pp. 451–471. DOI: 10.2307/41304213.
- Gladkova, A., A. Drozd, and S. Matsuoka (2016). “Analogy-Based Detection of Morphological and Semantic Relations with Word Embeddings: What Works and What Doesn’t”. In: *Proceedings of the NAACL Student Research Workshop*, pp. 8–15.
- Glorot, X., A. Bordes, and Y. Bengio (2011). “Domain Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach”. In: *Proceedings of the 28th International Conference on International Conference on Machine Learning*. Madison, WI, pp. 513–520.

Literatur VII

- Graves, A. et al. (2017). “Automated Curriculum Learning for Neural Networks”. In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, pp. 1311–1320.
- Gries, S. T. (2013). “50-Something Years of Work on Collocations”. In: *International Journal of Corpus Linguistics* 18.1, pp. 137–166.
- Gries, S. T. and D. Divjak (2009). “Behavioral Profiles: A Corpus-Based Approach to Cognitive Semantic Analysis”. In: *New directions in cognitive linguistics*, pp. 57–75.
- Gries, S. T. and P. Durrant (2019). “Analyzing Co-Occurrence Data”. In: Springer.
- Grigonytė, G. et al. (Aug. 2010). “Paraphrase Alignment for Synonym Evidence Discovery”. In: *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*. Beijing, China: Coling 2010 Organizing Committee, pp. 403–411.

Literatur VIII

- Gupta, A., G. Boleda, and S. Pado (Aug. 2018). “Instantiation”. In: *arXiv:1808.01662 [cs]*. arXiv: 1808.01662 [cs].
- Hamilton, W. L., J. Leskovec, and D. Jurafsky (2016). “Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change”. In: *arXiv preprint arXiv:1605.09096*.
- Harris, Z. S. (Aug. 1954). “Distributional Structure”. en. In: *Word* 10.2-3, pp. 146–162. DOI: 10.1080/00437956.1954.11659520.
- Haug, D. T. and M. Jøhndal (2008). “Creating a Parallel Treebank of the Old Indo-European Bible Translations [PROIEL]”. In: *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, pp. 27–34.
- Herbelot, A. and E. M. Vecchi (2015). “Building a Shared World: Mapping Distributional to Model-Theoretic Semantic Spaces”. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 22–32.

Literatur IX

- Hestness, J. et al. (Dec. 2017). “Deep Learning Scaling Is Predictable, Empirically”. In: *arXiv:1712.00409 [cs, stat]*. arXiv: 1712.00409 [cs, stat].
- Jiao, X. et al. (Dec. 2019). “TinyBERT: Distilling BERT for Natural Language Understanding”. In: *arXiv:1909.10351 [cs]*. arXiv: 1909.10351 [cs].
- Kaplan, J. et al. (Jan. 2020). “Scaling Laws for Neural Language Models”. In: *arXiv:2001.08361 [cs, stat]*. arXiv: 2001.08361 [cs, stat].
- Karakanta, A., J. Dehdari, and J. van Genabith (2018). “Neural Machine Translation for Low-Resource Languages without Parallel Corpora”. In: *Machine Translation* 32, pp. 167–189.
- Karan, M., J. Šnajder, and B. D. Bašić (2012). “Distributional Semantics Approach to Detecting Synonyms in Croatian Language”. In: *Information Society*, pp. 111–116.

Literatur X

- Lample, G. and A. Conneau (Jan. 2019). “Cross-Lingual Language Model Pretraining”. In: *arXiv:1901.07291 [cs]*. arXiv: 1901.07291 [cs].
- Lan, Z. et al. (Feb. 2020). “ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations”. In: *arXiv:1909.11942 [cs]*. arXiv: 1909.11942 [cs].
- Le, Q. V. and T. Mikolov (May 2014). “Distributed Representations of Sentences and Documents”. In: *arXiv:1405.4053 [cs]*. arXiv: 1405.4053 [cs].
- Lehecka, T. (2015). “Collocation and Colligation”. In: *Handbook of Pragmatics Online*. Ed. by J. Verschueren, J.-O. Östman, J. Blommaert, and C. Bulcaen. Benjamins.
- Levy, O. and Y. Goldberg (2014). “Neural Word Embedding as Implicit Matrix Factorization”. In: *Advances in Neural Information Processing Systems*, pp. 2177–2185.

Literatur XI

- Li, Y. et al. (2015). “Word Embedding Revisited: A New Representation Learning and Explicit Matrix Factorization Perspective”. In: *Twenty-Fourth International Joint Conference on Artificial Intelligence*, pp. 3650–3657.
- Luo, F. et al. (July 2018). “Incorporating Glosses into Neural Word Sense Disambiguation”. In: *arXiv:1805.08028 [cs]*. arXiv: 1805.08028 [cs].
- McGillivray, B. (2013). *Methods in Latin Computational Linguistics*. Brill.
- Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013a). “Efficient Estimation of Word Representations in Vector Space”. In: *arXiv preprint arXiv:1301.3781*.
- Mikolov, T. et al. (2013b). “Distributed Representations of Words and Phrases and Their Compositionality”. In: *Advances in Neural Information Processing Systems*, pp. 3111–3119.

Literatur XII

- Ono, M., M. Miwa, and Y. Sasaki (2015). “Word Embedding-Based Antonym Detection Using Thesauri and Distributional Information”. In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 984–989.
- Pennington, J., R. Socher, and C. Manning (Oct. 2014). “GloVe: Global Vectors for Word Representation”. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, pp. 1532–1543. DOI: 10.3115/v1/D14-1162.
- Peters, M. E. et al. (2018). “Deep Contextualized Word Representations”. In: *arXiv preprint arXiv:1802.05365*.
- Pinkal, M. (1993). “Semantik”. In: *Einführung in Die Künstliche Intelligenz*. Ed. by G. Görz. First, pp. 425–498.

Literatur XIII

- Punyakank, V., D. Roth, and W.-t. Yih (2008). “The Importance of Syntactic Parsing and Inference in Semantic Role Labeling”. In: *Computational Linguistics* 34.2, pp. 257–287.
- Rich, E. and K. Knight (Jan. 1991). *Artificial Intelligence*. Englisch. Subsequent. New York: McGraw-Hill Professional.
- Rogers, A., A. Drozd, and B. Li (2017). “The (Too Many) Problems of Analogical Reasoning with Word Vectors”. In: *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pp. 135–148.
- Sennrich, R., B. Haddow, and A. Birch (Aug. 2016). “Neural Machine Translation of Rare Words with Subword Units”. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, pp. 1715–1725. DOI: 10.18653/v1/P16-1162.

Literatur XIV

- Su, S.-Y., K.-L. Lo, Y.-T. Yeh, and Y.-N. Chen (Aug. 2018). “Natural Language Generation by Hierarchical Decoding with Linguistic Patterns”. In: *arXiv:1808.02747 [cs]*. arXiv: 1808.02747 [cs].
- Wiedemann, G., S. Remus, A. Chawla, and C. Biemann (Oct. 2019). “Does BERT Make Any Sense? Interpretable Word Sense Disambiguation with Contextualized Embeddings”. In: *arXiv:1909.10430 [cs]*. arXiv: 1909.10430 [cs].
- Wilkinson, M. D. et al. (Mar. 2016). “The FAIR Guiding Principles for Scientific Data Management and Stewardship”. en. In: *Scientific Data* 3, p. 160018. DOI: 10.1038/sdata.2016.18.