Introduction

HS Embeddings

Nils Reiter,
nils.reiter@uni-koeln.de

April 7, 2022
(Summer term 2022)

INSTITUT FÜR
DIGITAL HUMANITIES
UNIVERSITÄT ZU KÖLN

CRETA
CENTER FOR REFLECTED TEXT ANALYTICS

About Me

Figure: Nils (right)

Nils Reiter

- Diplom-Computerlinguist (Universität des Saarland)
- Promotion in Computerlinguistik (Uni Heidelberg, 2007-2013)
- Postdoc am IMS (Uni Stuttgart, 2014-2019)
- Jetzt: Professor für Digital Humanities/Sprachliche Informationsverarbeitung

- `https://nilsreiter.de`
  `nils.reiter@uni-koeln.de`

About Me

Research Interests

- Artistic/non-standard use of language (e.g., humor, art, metaphors, literature), why do we express things in a certain (individual!) way?
- Operationalisation of complex research questions and tasks
- Integration of quantitative/statistical research methods/results into hermeneutic research (e.g., interpretable machine learning)

→ 'Digital Humanities'

About Me

Research Interests

- Artistic/non-standard use of language (e.g., humor, art, metaphors, literature), why do we express things in a certain (individual!) way?
- Operationalisation of complex research questions and tasks
- Integration of quantitative/statistical research methods/results into hermeneutic research (e.g., interpretable machine learning)

→ 'Digital Humanities'

- ... also, I just like coding and team work

Sneak Peek

Embeddings: Representations of the meaning of words in vector spaces.

Sneak Peek

Embeddings: Representations of the meaning of words in vector spaces.
Meaning: The rest of the introduction

Sneak Peek

Embeddings: Representations of the meaning of words in vector spaces.
Meaning: The rest of the introduction

Vector space

- A collection of vectors of a fixed length
- Vector: An ordered set of numbers
  - Each number corresponds to one dimension
  - E. g.: $\vec{v} = (1, 3, 5, 6)$ (4 dimensions)
- Mathematical constructs

Sneak Peek

Embeddings: Representations of the meaning of words in vector spaces.
Meaning: The rest of the introduction

Vector space

- A collection of vectors of a fixed length
- Vector: An ordered set of numbers
  - Each number corresponds to one dimension
  - E. g.: $\vec{v} = (1, 3, 5, 6)$ (4 dimensions)
- Mathematical constructs
- Simply put: A word embedding is a function that assigns a vector to each word

Wer sind Sie?

- Vorstellungsrunde!
- Wer sind Sie?
- Warum sind Sie hier?
  - Hier: in diesem Kurs, in diesem Modul, in dieser Studienrichtung, in diesem Studiengang, ...?
- Was würden Sie machen, wenn Sie sich für etwas anderes entschieden hätten?

Section 2

Ablauf und Organisatorisches

Lernziele

- Lesen und Verstehen von (computerlinguistischer) Forschungsliteratur
- Aufbereitung eines Themas für andere
- Überblick über das Thema ‘Embeddings’ und seine Anwendungsmöglichkeiten

Modul: Verarbeitung von Textdaten

MA Informationsverarbeitung

Veranstaltung	Kontaktzeit	Selbststudium
Hauptseminar	30	60
Übung	30	60
Kolloquium	30	60
Modulprüfung	–	270

Tabelle: Lehrveranstaltungen im Modul

- 18 Leistungspunkte
- $30\,\%$ der Fachnote

Modul: Profilmodul Computerlinguistik (1C)

MA Linguistik

Veranstaltung	Kontaktzeit	Selbststudium
Hauptseminar	30	60
Projektseminar	30	60
Projektseminar	30	60
Modulprüfung	–	180

Tabelle: Lehrveranstaltungen im Modul

- 15 Leistungspunkte
- $40\,\%$ der Fachnote

Ablauf des Seminars

Zeitraum	Ref.	Inhalt
bis 05.05.	Prof.	Grundlagen zu Embeddings
12.05. – 30.06.	Stud.	Anwendungen von Embeddings
ab 07.07.	Prof.	Weiterführende Themen

Studentische Sitzungen

Ziel: Inhalt vermitteln, nicht: Papier referieren

- Mehr als zwei Wochen vorher: Literatur durcharbeiten, inklusive Forschungsstand
- Spätestens zwei Wochen vorher: Sprechstunde mit Dozent zu inhaltlichen Fragen zum Papier
- Spätestens eine Woche vorher: Konzept zum Referat per E-Mail abgeben. Das Konzept beinhaltet
  - die Gliederung/Struktur,
  - Aufgaben/Fragestellungen für Kleingruppen sowie erwartete Ergebnisse und
  - selbst erstellte Beispiele, die das gesagte verdeutlichen und konkretisieren
- Referat halten. Jedes Referat muss mind. eine Kleingruppenaktivität beinhalten
- Im Anschluss: Feedbackgespräch

Modulprüfungen

- Thema
  - Findung und Wahl: Ihre Aufgabe
  - Kann, muss aber nicht, etwas mit dem Seminar zu tun haben
  - Mit mir absprechen

Modulprüfungen

- Thema
  - Findung und Wahl: Ihre Aufgabe
  - Kann, muss aber nicht, etwas mit dem Seminar zu tun haben
  - Mit mir absprechen
- Praktischer Anteil: Offen.
  Beispiele: Experiment zur automatischen Identifikation eines Textphänomens, Annotationsexperiment, quantitativer Vergleich verschiedener Korpora, ...
- Am Ende: Hausarbeit von max 8 S. Länge
- Übung vor der Master-Arbeit

Modulprüfungen

- Thema
  - Findung und Wahl: Ihre Aufgabe
  - Kann, muss aber nicht, etwas mit dem Seminar zu tun haben
  - Mit mir absprechen
- Praktischer Anteil: Offen.
  Beispiele: Experiment zur automatischen Identifikation eines Textphänomens, Annotationsexperiment, quantitativer Vergleich verschiedener Korpora, ...
- Am Ende: Hausarbeit von max 8 S. Länge
- Übung vor der Master-Arbeit
- PDFs in Ilias mit Details

Fragen?

Section 3

Meaning

Let's talk about Meaning

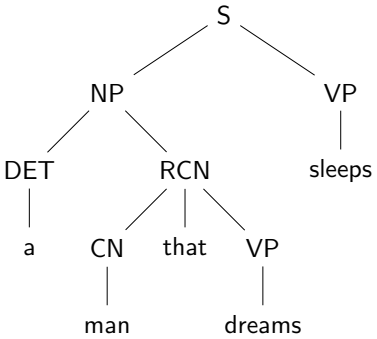
- Semantics: Linguistic study of meaning
- Linguistic expressions somehow mean something
  - Somehow: Not entirely clear how meaning works
  - Something: Different expressions mean different things

Let's talk about Meaning

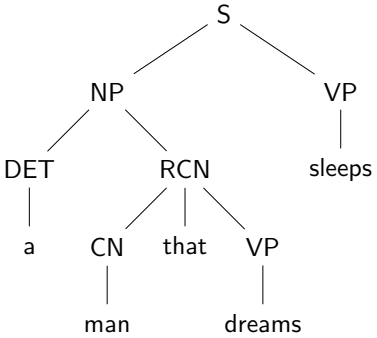
- Semantics: Linguistic study of meaning
- Linguistic expressions somehow mean something
  - Somehow: Not entirely clear how meaning works
  - Something: Different expressions mean different things
- Classical Semantics
  - Lexical: Meaning aspects stored in a lexicon of a language
  - Compositional: How lexicon entries are combined to meaning of larger structures (e. g., sentences)
    - Montague grammar

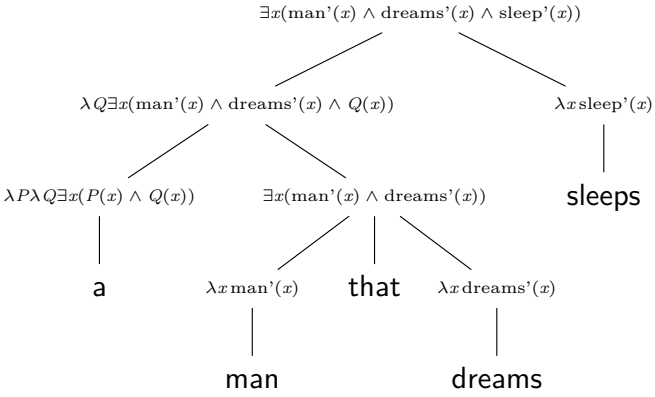
Montague (1970)

Montague Grammar (Montague, 1970)



Montague Grammar (Montague, 1970)





Meaning in Montague Grammar

- Lexicon
  - man $\leftrightarrow$ $\lambda x \, \mathrm{man'}(x)$
  - sleeps $\leftrightarrow$ $\lambda x \, \mathrm{sleep'}(x)$
  - ...

Meaning in Montague Grammar

- Lexicon
  - man $\leftrightarrow$ $\lambda x \, \mathrm{man'}(x)$
  - sleeps $\leftrightarrow$ $\lambda x \, \mathrm{sleep'}(x)$
  - ...
- Grammar
  - RCN $\rightarrow$ CN that VP $\leftrightarrow$ $\lambda x((CN\ x) \wedge (VP\ x))$
  - S $\rightarrow$ NP VP $\leftrightarrow$ $(NP\ VP)$ (functional application)
  - ...

Meaning as Truth Conditions

- Meaning: Truth conditions
  - What does it take to make a sentence true?
  - Model-theoretic interpretation:
    $\exists x(\mathrm{man'}(x) \wedge \mathrm{dreams'}(x) \wedge \mathrm{sleep'}(x))$ is true, iff the world contains something that is in the sets $\mathrm{man'}$, $\mathrm{dreams'}$, and $\mathrm{sleep'}$.

Is this sufficient? What is missing?

Discuss in small groups!

Is this sufficient? What is missing?

Lexical Relations

(Context-free) relations between lexical units (LU, stored in the lexicon of a language)

Lexical Relations

(Context-free) relations between lexical units (LU, stored in the lexicon of a language)

- Hyponymy/Hypernymy: General vs. specific
  - E. g.: Computer (hypernym) and MacBook Air (hyponym)

Lexical Relations

(Context-free) relations between lexical units (LU, stored in the lexicon of a language)

- Hyponymy/Hypernymy: General vs. specific
  - E. g.: Computer (hypernym) and MacBook Air (hyponym)
- Synonymy: Different LUs that have the same meaning (or: roughly the same)
  - Not every LU has synonyms
  - E. g.: happy, joyful, glad

Lexical Relations

(Context-free) relations between lexical units (LU, stored in the lexicon of a language)

- Hyponymy/Hypernymy: General vs. specific
  - E. g.: Computer (hypernym) and MacBook Air (hyponym)
- Synonymy: Different LUs that have the same meaning (or: roughly the same)
  - Not every LU has synonyms
  - E. g.: happy, joyful, glad
- Antonymy: LUs with opposite meanings
  - Not every LU has an antonym
  - E. g.: cold vs. warm; dead vs. alive

Lexical Relations

(Context-free) relations between lexical units (LU, stored in the lexicon of a language)

- Hyponymy/Hypernymy: General vs. specific
  - E. g.: Computer (hypernym) and MacBook Air (hyponym)
- Synonymy: Different LUs that have the same meaning (or: roughly the same)
  - Not every LU has synonyms
  - E. g.: happy, joyful, glad
- Antonymy: LUs with opposite meanings
  - Not every LU has an antonym
  - E. g.: cold vs. warm; dead vs. alive
- Homonymy (a.k.a. semantically ambiguous): A LU that has multiple meanings
  - E. g.: bank (financial institution) vs. bank (of a river)

Ludwig Wittgenstein: Philosophische Untersuchungen (1953)

1. Über etablierte Sprachphilosophie:
   Jedes Wort hat eine Bedeutung. Diese Bedeutung ist dem Wort zugeordnet. Sie ist der Gegenstand, für den das Wort steht.

Ludwig Wittgenstein: Philosophische Untersuchungen (1953)

1. Über etablierte Sprachphilosophie:

   Jedes Wort hat eine Bedeutung. Diese Bedeutung ist dem Wort zugeordnet. Sie ist der Gegenstand, für den das Wort steht.

2.

   Jener philosophische Begriff der Bedeutung ist in einer primitiven Vorstellung von der Art und Weise, wie die Sprache funktioniert, zu Hause. Man kann aber auch sagen, es sei die Vorstellung einer primitiveren Sprache, als der unsern.

Ludwig Wittgenstein: Philosophische Untersuchungen (1953)

1. Über etablierte Sprachphilosophie:

   Jedes Wort hat eine Bedeutung. Diese Bedeutung ist dem Wort zugeordnet. Sie ist der Gegenstand, für den das Wort steht.

2.

   Jener philosophische Begriff der Bedeutung ist in einer primitiven Vorstellung von der Art und Weise, wie die Sprache funktioniert, zu Hause. Man kann aber auch sagen, es sei die Vorstellung einer primitiveren Sprache, als der unsern.

43.

   *Man kann für eine **große** Klasse von Fällen der Benützung des Wortes **Bedeutung** – wenn auch nicht für **alle** Fälle seiner Benützung – dieses Wort so erklären: Die Bedeutung eines Wortes ist sein Gebrauch in der Sprache.*

"The meaning of a word is its use in the language"

- No logical formulas to represent word meaning
- Represent "how the word is used by actual people in speaking and understanding" (Jurafsky/Martin, 2019)

"The meaning of a word is its use in the language"

The distribution of an element will be understood as the sum of all its environments. An environment of an element A is an existing array of its co-occurrents, i. e. the other elements, each in a particular position, with which A occurs to yield an utterance. (Harris, 1954, 146)

it is possible to divide [...] any flow of speech into parts, in such a way that we can find some regularities in the occurrence of one part relative to others in the flow of speech. (Harris, 1954, 158)

"The meaning of a word is its use in the language"

The distribution of an element will be understood as the sum of all its environments. An environment of an element A is an existing array of its co-occurrents, i. e. the other elements, each in a particular position, with which A occurs to yield an utterance. (Harris, 1954, 146)

it is possible to divide [...] any flow of speech into parts, in such a way that we can find some regularities in the occurrence of one part relative to others in the flow of speech. (Harris, 1954, 158)

You shall know a word by the company it keeps (Firth, 1957, 11)

Distributional Semantics

(1) Ongchoi is delicious sauteed with garlic.

(2) Ongchoi is superb over rice.

(3) ... ongchoi leaves with salty sauces ...

What is ongchoi?

Section 4

Summary

Summary

- Semantics: Study of meaning (of language utterances)

Summary

- Semantics: Study of meaning (of language utterances)
- Different forms of semantics
- Classical: Lexical and compositional semantics
  - E. g., Montague grammar

Summary

- Semantics: Study of meaning (of language utterances)
- Different forms of semantics
- Classical: Lexical and compositional semantics
  - E. g., Montague grammar
- Distributional semantics
  - Inspired by Wittgenstein, Harris, ...
  - Look at distributions of language use in natural language

Summary

- Semantics: Study of meaning (of language utterances)
- Different forms of semantics
- Classical: Lexical and compositional semantics
  - E. g., Montague grammar
- Distributional semantics
  - Inspired by Wittgenstein, Harris, ...
  - Look at distributions of language use in natural language
- Sneak peak
  - Distributions = co-occurrence counts
  - Word vector = co-occurrence counts with other words

References I

- Firth, John Rupert (1957). *A synopsis of linguistic theory, 1930-1955*.
- Harris, Zellig S. (1954). "Distributional Structure". In: *WORD* 10.2-3, pp. 146–162.
- Jurafsky, Dan/James H. Martin (2019). *Speech and Language Processing*. 3rd ed. Draft of October 16, 2019. Prentice Hall.
- Montague, Richard (1970). "Universal grammar". In: *Theoria* 36.3, pp. 373–398. DOI: 10.1111/j.1755-2567.1970.tb00434.x.