UNIVERSITÄT
ZU KÖLN

Einführung in die Informationsverarbeitung

Nils Reiter

October 12, 2023

INSTITUT FÜR
DIGITAL HUMANITIES
UNIVERSITÄT ZU KÖLN

Sprachliche Informationsverarbeitung

Course topics

- Overview: Language processing
- Linguistic areas and phenomena
- Corpus linguistics and statistics
- Annotation workflow
- Machine learning

Sprachliche Informationsverarbeitung

- Information often encoded in language
    - E.g., on this slide or in this lecture

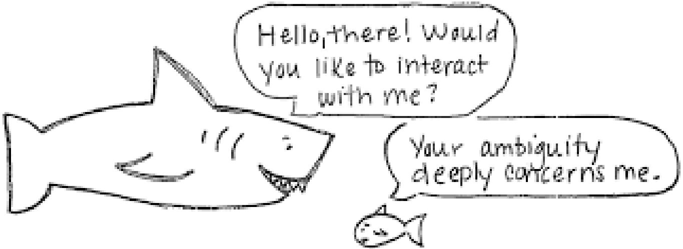
Sprachliche Informationsverarbeitung

- Information often encoded in language
  - E.g., on this slide or in this lecture
- Harvesting language automatically is difficult
  - Language is in many ways ambiguous
  - Meaning of words can change (Mouse as animal vs. as input device)
  - Language rules are evolving

Sprachliche Informationsverarbeitung

- Information often encoded in language
  - E.g., on this slide or in this lecture
- Harvesting language automatically is difficult
  - Language is in many ways ambiguous
  - Meaning of words can change (Mouse as animal vs. as input device)
  - Language rules are evolving
- Text production increases
  - Average student, average day: 15k (spoken) words Mehl et al. (2007)
  - Average US-American: 94 text messages per day TextRequest

  ⇒ There is a gigantic amount of words out there!


Hello, there! Would you like to interact with me?
Your ambiguity deeply concerns me.

Language Ambiguity

Language Ambiguity

- A sentence is ambiguous: There are multiple possible readings/meanings

Language Ambiguity

- A sentence is ambiguous: There are multiple possible readings/meanings
- Fundamental property of natural language
- Often basis for humor
- Takes place on all language levels
    - Sentences can be ambiguous
        - What is their syntactic structure?
    - Words can be ambiguous
        - What is their morphological structure?
    - Words can be ambiguous in their context
        - To which character does a pronoun refer to?

Language Ambiguity

Examples

Der Jäger traf den Mann mit dem Gewehr.

Language Ambiguity

Examples

Landesmusikdirektor:in

Language Ambiguity

Examples

Landesmusikdirektor:in

Musikdirektor:in des Landes

Direktor:in für Landesmusik

Language Ambiguity

Examples

Maria hat Petra beim Einkaufen getroffen. Sie hat ihr Geld geliehen.

Language Ambiguity

Examples

Maria ging zur Bank.

Language Ambiguity

Examples

Maria ging zur Bank und setzte sich hin.

Language Ambiguity

Examples

Maria ging zur Bank und raubte sie aus.

Language Ambiguity

Examples

Language Ambiguity

Examples

hubert hat dort liebe genossen.

Language Ambiguity

Examples

hubert hat dort liebe genossen.

Hubert hat dort Liebe genossen.

Hubert hat dort liebe Genossen.

Language Ambiguity

Examples

Time flies like an arrow.

Machine learning vs. Rule-based

Example

Grammar rules:

- A nominal phrase (NP) contains a determiner and a noun
  - “the dog”/“Der Hund” is a noun phrase

Machine learning vs. Rule-based

Example

Grammar rules:

- A nominal phrase (NP) contains a determiner and a noun
  - "the dog"/"Der Hund" is a noun phrase
- Subject and verb agree in number
  - "the dog barks" / "Der Hund bellt." is grammatical
  - *"the dog bark" / *"Die Hunde bellt." is not (because NP and verb have different numbers)

Machine learning vs. Rule-based

Example

Grammar rules:

- A nominal phrase (NP) contains a determiner and a noun
  - “the dog”/“Der Hund” is a noun phrase
- Subject and verb agree in number
  - “the dog barks” / “Der Hund bellt.” is grammatical
  - *“the dog bark” / *“Die Hunde bellt.” is not (because NP and verb have different numbers)

Two options for processing language

- Rule-based systems
  - Write programs that directly implements grammar rules

Machine learning vs. Rule-based

Example

Grammar rules:

- A nominal phrase (NP) contains a determiner and a noun
  - “the dog”/“Der Hund” is a noun phrase
- Subject and verb agree in number
  - “the dog barks” / “Der Hund bellt.” is grammatical
  - *“the dog bark” / *“Die Hunde bellt.” is not (because NP and verb have different numbers)

Two options for processing language

- Rule-based systems
  - Write programs that directly implements grammar rules
- Machine learning
  - Write programs that learn grammar rules from data

Why?

- Grammar rules are documented, why not just implement them in a program?
    - Rule systems

Why?

- Grammar rules are documented, why not just implement them in a program?
    - Rule systems
- Language is more productive and creative
- Grammar rules are not complete
- Lexicons are even farther from being complete – and likely will be, forever

Brief history of Computational Linguistics I

- 1950s: DARPA Projects to automatically translate Russian into English
- 1957/65: Linguistics shifts focus from describing to generating Chomsky (1957, 1965)
- 1959: Theo Lutz for the first time generates a German poem with a computer Lutz (1959)
- 1962: Foundation of the "Association for Machine Translation and Computational Linguistics", 1968 renamed to "Association for Computational Linguistics (ACL)"
- 1966, ALPAC report: MT more expensive, less accurate and slower than human translation ALPAC (1966)
- 1968: Foundation of SYSTRAN, first MT company
- 1975: European commission uses SYSTRAN software (first use of MT on EU level)

Brief history of Computational Linguistics II

- 1984: First corpus-based commercial MT system Nagao (1984)
- 1992: Study programs established in Germany (Universities Saarbrücken/Stuttgart)
- 2011: IBM Watson beats two humans in Jeopardy / Apple Siri launched https://www.youtube.com/watch?v=WFR3lOm_xhE
- 2013: Word embeddings (e.g., word2vec) Mikolov et al. (2013)
- 2017: Launch of the DeepL Translator
- 2018: Transformer models: BERT Devlin et al. (2019)
- 2022: Publicly usable transformer model ChatGPT https://chat.openai.com

Course Topics

- Linguistic areas and phenomena
- Corpus linguistics and statistics
- Annotation and its verification
- Machine learning

References I

- ALPAC (1966). *Language and Machines. Computers in Translation and Linguistics.* Tech. rep. National Research Council.
- Chomsky, Noam (1957). *Syntactic Structures*. Mouton De Gruyter.
- — (1965). *Aspects of the theory of syntax*. MIT Press.
- Devlin, Jacob/Ming-Wei Chang/Kenton Lee/Kristina Toutanova (2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423.
- Lutz, Theo (1959). “Stochastische Texte”. In: *augenblick* 4, pp. 3–9. URL: https://www.netzliteratur.net/lutz%5C_schule.htm.

References II

- Mehl, Matthias R./Simine Vazire/Nairán Ramírez-Esparza/Richard B. Slatcher/James W. Pennebaker (2007). "Are Women Really More Talkative Than Men?" In: *Science* 317. DOI: 10.1126/science.1139940.
- Mikolov, Tomáš/Kai Chen/Greg Corrado/Jeffrey Dean (2013). "Efficient Estimation of Word Representations in Vector Space". In: *arXiv cs.CL*. URL: https://arxiv.org/pdf/1301.3781.pdf.
- Nagao, Makoto (1984). "A Framework of a Mechanical Translation between Japanese and English by Analogy Principle". In: *Proc. of the International NATO Symposium on Artificial and Human Intelligence*. Lyon, France: Elsevier North-Holland, Inc., pp. 173–180. ISBN: 0444865454.