

Information Extraction

Referatsthemen

CIS, LMU München
Winter Semester 2016-2017

Dr. Alexander Fraser, CIS

Information Extraction – Reminder

- Vorlesung
 - Learn the basics of Information Extraction (IE), **Klausur – only on the Vorlesung!**
- Seminar
 - Deeper understanding of IE topics
 - Each student who wants a Schein will have to make a presentation on IE
 - 25 minutes (powerpoint, LaTeX, Mac)
 - THESE NUMBERS MAY CHANGE AS I MAKE THE SCHEDULE!
- Hausarbeit
 - 6 page "Ausarbeitung" (an essay/prose version of the material in the slides), **due 3 weeks after the Referat**
 - Optionally: bonus points from practical exercises

Topics

- Topic will be presented in roughly the same order as the related topics are discussed in the Vorlesung
- Most of the topics require you to do a literature search
 - There will usually be one article (or maybe two) which you find is the key source
 - For some topics, a suggestion will be made on the slide
 - If these sources are not standard peer-reviewed scientific articles, **YOU MUST SEND ME AN EMAIL 2 WEEKS BEFORE YOUR REFERAT** to ask permission
- There are a few projects involving programming
- I am also open to topic suggestions suggested by you, send me an email

Referat

- Tentatively (MAY CHANGE!):
 - 25 minutes plus about 15 minutes for discussion
- Start with what the problem is, and why it is interesting to solve it (motivation!)
 - It is often useful to present an example and refer to it several times
- Then go into the details
- If appropriate for your topic, do an analysis
 - Don't forget to address the disadvantages of the approach as well as the advantages
 - Be aware that advantages tend to be what the original authors focused on!
- **List references and recommend further reading**
- **Number your slides (useful in discussion)**
- **Have a conclusion slide!**
- **IMPORTANT: if your topic is repeated from a previous year's seminar, explicitly (but briefly) say what was done there and how your presentation is different!**

Languages

- I recommend:
- If you do the slides in English, then presentation in English (and Hausarbeit in English)
- If you do the slides in German, then presentation in German (and Hausarbeit in German)
- Additional option (not recommended):
 - English slides, German presentation, English Hausarbeit
 - Very poor idea for non-native speakers of German (you will get tired by the end of the discussion because English and German interfere)

References

- Please use a standard bibliographic format for your references
- This includes authors, date, title, venue, like this:
- Academic Journal
 - Alexander Fraser, Helmut Schmid, Richard Farkas, Renjing Wang, Hinrich Schuetze (2013). Knowledge Sources for Constituent Parsing of German, a Morphologically Rich and Less-Configurational Language. *Computational Linguistics*, 39(1), pages 57-85.
- Academic Conference
 - Alexander Fraser, Marion Weller, Aoife Cahill, Fabienne Cap (2012). Modeling Inflection and Word-Formation in SMT. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 664-674, Avignon, France, April.

References II

- In the Hausarbeit, use **inline** citations:
 - "As shown by Fraser et al. (2012), the moon does not consist of cheese"
 - "We build upon previous work (Fraser and Marcu 2007; Fraser et al. 2012) by ..."
 - Sometimes it is also appropriate to include a page number (and you **must** include a page number for a quote or graphic)
- Please do not use numbered citations like:
 - "As shown by [1], ..."
 - Numbered citations are useful to save space, otherwise quite annoying

References III

- If you use graphics (or quotes) from a research paper, MAKE SURE THESE ARE CITED ON THE *SAME SLIDE* IN YOUR PRESENTATION!
 - These should be cited in the Hausarbeit in the caption of the graphic
 - Please include a page number so I can find the graphic quickly
- Web pages should also use a standard bibliographic format, particularly including the date when they were downloaded
- I am not allowing Wikipedia as a primary source
 - I no longer believe that Wikipedia is reliable, for most articles there is simply not enough review (mistakes, PR agencies trying to sell particular ideas anonymously, etc.)
 - Wikipedia can be useful for background, but please don't cite Wikipedia pages!
- You also cannot use student work (not peer-reviewed by people with PhDs) as a primary source
 - If in doubt, email me!

Administravia - I

- The Referatsthemen will be presented in both seminars this week
- You will then send me an email with your preferences *starting at 14:00 Wednesday, November 2nd*
 - Do not forget to say WHICH DAY OF THE WEEK (Wed or Thurs)
 - Emails will be processed in the order received
 - Emails received before 14:00, even one minute before, will be processed later, this is the only fair way to allocate topics
 - You can specify multiple topics (ranked) if you want
 - I might counter-propose a similar topic if there is a topic that you particularly want (but I might not be able to do this, don't count on it!)

Administravia - II

- You can take a look at topics from the previous year (you should know the password)
- First seminar topics are already in two weeks!
 - **For this reason, for the first two topics ONLY**, please send me an email starting at Friday 10:00
- You can look at the seminar web page as I update it, click the refresh button in your browser due to possible caching problems

Administravia - III

- Tutor: Fabian Dreer
- Email: dreer@cip.ifi.lmu.de

- Wednesday 02.11 and Thursday 03.11 Fabian Dreer and I will hold an exercise on manually written rules for extraction in the Seminar
 - People only in the VL are also invited
 - See the Seminar web page for location, will be posted soon

Information Extraction

Information Extraction (IE) is the process of extracting structured information from unstructured machine-readable documents

and beyond

Ontological Information Extraction

Fact Extraction

Instance Extraction

Named Entity Recognition

Tokenization & Normalization

Source Selection

05/01/67



1967-05-01

...married Elvis
on 1967-05-01

Elvis Presley	Singer
Angela Merkel	Politician

- The first four topics must be in English (this is marked on each slide)
- Two common pitfalls:
 - Please provide the motivation for your topic!
 - PLEASE DO NOT FORGET SLIDE NUMBERS!

History of IE

- TOPIC: MUC, ACE and TAC (Knowledgebase Population Track)
 - These workshops worked on Information Extraction, funded by US but a large variety of people participated
 - Discuss problems solved, motivations and techniques
 - Survey the literature
- **MUST BE IN ENGLISH**

(Tsuyoshi Okita)

Named Entity Recognition – Entity Classes

- TOPIC: fine-grained open classes of named entities
 - Survey the proposed schemes of fine-grained open classes, such as BBN's classes used for question answering
 - Discuss the advantages and disadvantages of the schemes
 - Discuss also the difficulty of human annotation – can humans annotate these classes reliably?
 - How well do classification systems work with these fine grained classes?

Rule-based IE vs. Statistical

- TOPIC: Rule-based IE (dominant in industry) vs. Statistical IE (dominant in academia)
 - Discuss the academic history of IE
 - What is the general view in academia towards rule-based IE?
 - How is statistical IE viewed in industry?
 - Time allowing: hybrid

NER – Stanford Toolkit

- TOPIC: Stanford NER Toolkit applied to TED2013
 - Apply the Stanford NER Toolkit to the TED2013 corpus (taken from the OPUS corpus), and compare the output on English and German
 - How does the model work (differentiate between English and German)?
 - What are the differences between the English and German annotations of parallel sentences, where do the models fail?

NER – OpenNLP Toolkit

- TOPIC: OpenNLP NER Toolkit applied to TED2013
 - Apply the OpenNLP NER Toolkit to the TED2013 corpus (taken from the OPUS corpus), and compare the output on English and German
 - How does the model work (differentiate between English and German)?
 - What are the differences between the English and German annotations of parallel sentences, where do the models fail?

NER – Twitter

- TOPIC: Named Entity Recognition of Entities in Twitter
 - There has recently been a lot of interest in annotating Twitter
 - Which set of classes is annotated? What is used as supervised training material, how is it adapted from non-Twitter training sets?
 - What are the peculiarities of working on 140 character tweets rather than longer articles?

NER – BIO Domain

- TOPIC: Named Entity Recognition of Biological Entities
 - Present a specific named entity recognition problem from the biology domain
 - Which set of classes is annotated? What is used as supervised training material?
 - What are the difficulties of this domain vs. problems like extraction of company mergers which have been studied longer?

Instance Extraction – Coreference

- TOPIC: surveying the literature on Coreference
 - How do existing pipelines work? What are the differences?
 - What gold standard data is available for testing systems?
 - What types of coreference are detected?
 - How do the models work?
 - What sort of results does one get?
 - What are the open problems?

Instance Extraction - Coref with Stanford

- TOPIC: Applying the Stanford Coreference Pipeline to TED2013 (from the OPUS corpus)
 - Apply the Stanford Coreference Pipeline to English TED2013 data
 - Discuss the general pipeline and how it works
 - What entities in TED2013 does it annotate well, and less well?
 - Can this information be used to translate English "it" to German?

(Fabienne Braune and Matthias Huck)

Choosing a topic

- Any questions?
- I will put these slides on the seminar page later today
- Please email me with your choice of topic, starting at *14:00 Wednesday November 2nd*
- **You must also say which day you want to present** (Wed, Thurs, or both days possible)!
 - If you are emailing later, check the seminar web page first to see if the topic is already taken!

- Thank you for your attention!