Information Extraction
Referatsthemen
CIS, LMU München
Winter Semester 2015-2016
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 (this is optional!)
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
  • 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
- 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)
    • (Academic Conference)
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
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
  - After looking into it, 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.)

You also cannot use student work (not peer-reviewed) as a primary source
Information Extraction (IE) is the process of extracting structured information from unstructured machine-readable documents.

- **Source Selection**
- **Tokenization & Normalization**
  - 05/01/67 → 1967-05-01
- **Named Entity Recognition**
  - ...married Elvis on 1967-05-01

**Ontological Information Extraction**

**Fact Extraction**

**Instance Extraction**

- **Elvis Presley**  Singer
- **Angela Merkel**  Politician

Slide from Suchanek
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
Source Selection

• TOPIC: Focused web crawling
  • Why use focused web crawling?
  • How do focused web crawlers work?
  • What are the benefits and disadvantages of focused web crawling?
• Example toolkits:
  • Python: scrapy
  • Perl: WWW::Mechanize
Source Selection

• TOPIC: Wrappers
  • Wrappers are used to extract tuples (database entries) from structured web sites
  • Discuss the different ways to create wrappers
    • Advantages and disadvantages
    • How do wrappers deal with changing websites?
  • Give some examples of different wrapper creation software packages and discuss their pros and cons
Rule-based Named Entity Recognition

• TOPIC: Parsing Resumes
  • Why is it important to parse resumes and how is the information used?
  • What sort of entities occur in resumes and how are they detected?
  • How are resumes parsed using rules? How is the problem structured, what is the overall approach?
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?
Named Entity Recognition – Training Data

• TOPIC: Crowd-sourcing with Amazon Mechanical Turk (AMT)
  • AMT's motto: artificial intelligence
  • Using human annotators to get quick (but low quality) annotations
  • What are the pros and cons of this approach?
  • Present how NER data is collected using AMT
  • How well do NER systems perform when trained on this data?
Named Entity Recognition - Supervision

• TOPIC: Lightly Supervised Named Entity Recognition
  • Starting from a few examples ("seed examples"), how do you automatically build a named entity classifier?
    • This is sometimes referred to as "bootstrapping"
  • What the problems with this approach – how do you block the process from generalizing too much?
  • Analyze the pros and cons of this approach
Named Entity Recognition - Supervision

• TOPIC: Distant supervision for NER
  • Related to the bootstrapping idea – but here we are using information annotated for a different purpose
  • How can distant supervision solve the knowledge bottleneck for NER?
  • What are the advantages and disadvantages of this approach?
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?
Classification-based Citation Parsing

• TOPIC: parsing citations using classifiers
  • How is the citation parsing problem formulated using classifiers?
  • What sort of information is available?
  • What does the training data look like?
  • What sorts of downstream applications are based on citation parsing?
NER – Stanford Toolkit

• **TOPIC: Stanford NER Toolkit applied to EMEA**
  • Apply the Stanford NER Toolkit to the EMEA 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 EMEA
  • Apply the OpenNLP NER Toolkit to the EMEA 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 – Domain Adaptation

• TOPIC: Domain adaptation and failure to adapt
  • What is the problem of domain adaptation?
  • How is it addressed in statistical classification approaches to NER?
  • How well does it work?
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 EMEA (from the OPUS corpus)
  • Apply the Stanford Coreference Pipeline to English EMEA data
  • Discuss the general pipeline and how it works
  • What entities in EMEA does it annotate well, and less well?
  • Can this information be used to translate English "it" to German?
Event Extraction – Epidemics

• TOPIC: Extracting Information about epidemics (for example, from ProMED-mail)
  • How do existing pipelines work?
  • What gold standard data is available for testing systems?
  • What are the entities detected?
  • How is the information aggregated?
  • How can the information be used?
Event Extraction – Disasters in Social Media

• TOPIC: Extracting Information during a disaster from social media (e.g., Twitter)
  • What sorts of real-time information extraction can be done using social media?
  • What are the entities detected?
  • How is the information aggregated?
  • How can the information be used?
IE for multilingual applications

• TOPIC: Evaluating automatically extracted bilingual lexica
  • The problem of word alignment is the task of finding terms which are translations of each other given their context in parallel corpora
  • How can these be compiled into bilingual lexica?
  • How can these lexica be evaluated? What are the critical sources of knowledge for this evaluation?
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 *19:00 Thursday October 29th*

• **You must also say which day you want to present** (Wed, Thurs, or both days possible)!
  • If you are emailing later, check the seminar page first to see if the topic is already taken!
• Thank you for your attention!