Information Extraction
Referatsthemen
CIS, LMU München
Winter Semester 2021-2022
Prof. 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
    - New: 3 (sub-)presentations on a single topic, each are 10 minutes (LaTeX, PowerPoint, Keynote)
  - THIS MAY CHANGE A LITTLE AS I MAKE THE SCHEDULE!
    - If so, I will tell you this next week in the Vorlesung

- Hausarbeit
  - 4 page "Ausarbeitung" (an essay/prose version of the material in the slides), due 3 weeks after the Referat
  - One Hausarbeit per student, submitted separately, per email!
Why this Seminar (not an Übung)?

• Develop competence in carrying out a literature review, writing and presentation
• Has similarities to the Bachelorarbeit you will do next semester
• Good practice for the Masters, there are many seminars
• Note: Getting a good grade here will be useful for the 2.50 average requirement for the Masters, which is now in effect
• Learn by observing what other students do well, but also not so well
Topics

• Topic will be presented in roughly the same order as the related topics are discussed in the Vorlesung

• To understand the topics fully requires you to do a literature search
  • There will usually be one article (or maybe two) which you find is the key source for your presentation
    • For some topics, a suggestion will be made on the slide
  • If the sources you use are not standard peer-reviewed scientific articles, YOU MUST SEND ME AN EMAIL 2 WEEKS BEFORE YOUR REFERAT to ask permission
  • If a paper is behind a paywall, try to use the E-Media service of the LMU library (using your LMU Kennung):
    • [https://www.ub.uni-muenchen.de/e-medien-der-ub/index.html](https://www.ub.uni-muenchen.de/e-medien-der-ub/index.html)

• All students will present at least one paper (!)
Referat

• Tentatively (MAY CHANGE!):
  • 3 presentations, each is 10 minutes. 15 minutes for the advisor to ask questions, a few more minutes for discussion
• The first student will present the problem, the motivation and a single paper
  • The first presentation starts with what the overall 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
• The second student will present one or two papers on different approaches to the problem
• The third student will present the most recent paper and an analysis (brief comparison of the different approaches) and a conclusion
  • Don't forget to address the disadvantages of the approaches as well as the advantages
  • Be aware that advantages tend to be what the original authors focused on!
Important tips

• List references and recommend further reading!
• Number your slides (useful in discussion)!

• The three students working on a single topic need to coordinate! Have one outline clearly indicating where the transitions between students are
  • Show this at the start of each of the sub-presentations
• IMPORTANT: practice the talk in the group, and give each other feedback to improve the talk
Language

• 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)
• You must specify the presentation language when you specify topics, I will use this in scheduling the topics
• Each set of three topics is in a single language!
References I

• Please use a standard bibliographic format for your references

• This includes authors, date, title, venue, like this:

• Academic Journal

• Academic Conference
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:

- DO NOT USE: "As shown by [1], ..."
- DO NOT USE: footnotes containing the citations
- 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

• Please send me an email with your preferences
  • Starting at 18:00 on *Friday*
  • The email sender *must* CC the other two students!
  • Please say which seminar (weekday) you are in (and your names)
  • Specify which language you will present in
  • Emails will be processed in the order received
  • Emails received before 18:00, even one minute before, will be processed later, this is the only fair way to allocate topics
  • You can specify multiple topics (ranked)
• Last topics assigned on Wednesday next week, this is the deadline!
Administravia II

• You can look at the seminar web page as I update it, click the refresh button in your browser due to possible caching problems
• First seminar topics are already in just two weeks!
Administravia III – Corona

• I expect that at least the first few presentations will be live
• Within the next few weeks, Bavaria will reach the "Red Krankenhausampel"
  • At this point, I expect we may switch to zoom, but LMU has not decided this yet
Administraviva IV – Live version

• Corona rules
  • You need to wear a mask when not speaking, FFP2 will be required soon ("Gelbe Krankenhausampel")
  • When speaking without a mask you need to be well over 1.5 meters away from everyone else, otherwise use a mask

• Please check that all laptops being used can actually project with the projector in the seminar room

• Rehearse the talk so that you know it really ends after 10 minutes each. I will cut you off shortly after this time limit!

• PLEASE DO NOT FORGET THE SLIDE NUMBERS!
Administrivia IV – Zoom version

• Please check that zoom presentations are working for you as a group! Make sure that your cameras and audio are working.
• Rehearse the talk so that you know it really ends after 10 minutes each. I will cut you off shortly after this time limit!
• PLEASE DO NOT FORGET THE SLIDE NUMBERS!
• Questions?
Information Extraction

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

- **Source Selection**
- **Tokenization & Normalization**
- **Named Entity Recognition**
- **Instance Extraction**
- **Fact Extraction**
- **Ontological Information Extraction**

<table>
<thead>
<tr>
<th>Person Name</th>
<th>Person Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elvis Presley</td>
<td>musician</td>
</tr>
<tr>
<td>Angela Merkel</td>
<td>politician</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Relation</th>
<th>Entity1</th>
<th>Entity2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Married</td>
<td>Elvis Presley</td>
<td>Priscilla Beaulieu</td>
</tr>
<tr>
<td>CEO</td>
<td>Tim Cook</td>
<td>Apple</td>
</tr>
</tbody>
</table>

05/01/67 → 1967-05-01

... married Elvis on 1967-05-01

And Beyond!

Tip of the hat: Suchanek
• Some of my topics must be in English

• Two common pitfalls:
  • Please provide the motivation for your topic!
  • PLEASE DO NOT FORGET SLIDE NUMBERS!
History of IE

• TOPIC: History of IE, shared tasks
• Three different workshop series:
  • MUC
  • ACE
  • TAC
• These workshops worked on Information Extraction, funded by US but a large variety of research groups participated
• Discuss problems solved, motivations and techniques
• Survey the literature
• Present the specific 2020 task and the best system:
  • Recognizing Ultra Fine-grained Entities (RUFES) 2020
• Optionally present alternative systems

MUST BE IN ENGLISH
Named Entity Recognition – Entity Classes

• TOPIC: fine-grained open classes of named entities
  • Survey the proposed schemes of fine-grained open classes:
    • 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?

• MUST BE IN ENGLISH
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?

• PAPER: please select a 2020 or 2021 paper as the final primary source, use the citation chain to find two or three previous papers
Creating Training Data with Weak Supervision for Relation Extraction

• TOPIC: using rules instead of hand-labeling training data for relation extraction
  • All machine learning based systems are heavily dependent on large training data
  • But domain experts can often write rules effectively that capture important generalizations
  • Can we use these rules to augment supervised relation extraction systems?

• Recommended Papers:
  • Braden Hancock, Martin Bringmann, Paroma Varma, Percy Liang, Stephanie Wang, and Christopher Ré (2018). Training Classifiers with Natural Language Explanations. ACL, pages 1884-1895.
Coreference

- Coreference systems have made many improvements recently.
- This topic will discuss the basic problem of coreference, then present several papers on recent work on coreference systems.
- Suggested third paper (optional):
- Vladimir Dobrovolskii (2021). Word-Level Coreference Resolution. EMNLP
Open IE Systems

- TOPIC: doing relation extraction with no templates and no pre-defined entity types. Just read the web and build a knowledge base.
  - This is an exciting area right now, real advances are being made
  - How is this done? Which machine learning techniques are used? How is the system initialized?
  - How can we evaluate such systems?
- Recommended third paper:
• (Viktor Hangya, Jindrich Libovicky, Katharina Hämmerl, Alexandra Chronopoulou)
Choosing a topic

- Any questions?
- I will put these slides on the seminar page later today
- Please email me with your choice of topics (FOR ALL TOPICS!), starting at *18:00* Friday (it may take me some time to get back to you, I am at the EMNLP conference next week)
  - Do not forget to include the presentation language (and your names!)
  - Do not forget to CC your co-presenters
- If you are emailing later, check the seminar web page first to see if the topic is already taken!
• Thank you for your attention!