Information Extraction
Topics

CIS, LMU München
Winter Semester 2022-2023

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 9 minutes (LaTeX, PowerPoint, Keynote)
  • THIS MAY CHANGE A LITTLE AS I MAKE THE SCHEDULE!
    • If so, I will tell you this next time in the Vorlesung

• Hausarbeit
  • 4 page "Auszarbeitung" (an essay/prose version of the material in the slides), **due 3 weeks after the Referat**
  • **One Hausarbeit per student, submitted separately, per email!**
Administravia I

• Please send me an email with your preferences
  • Starting at 18:00 on *Thursday* (tomorrow!)
  • The email sender *must* CC the other two students!
  • Please say 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 in three weeks
Administravia III

- 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 9 minutes each. I will cut you off shortly after this time limit!
- PLEASE DO NOT FORGET THE SLIDE NUMBERS!
• Questions?
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>
• 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:
  - Epidemic Question Answering (EPIC-QA) 2020
  - Optionally present alternative systems
- MUST BE IN ENGLISH
Named Entity Recognition – Entity Classes

• TOPIC: fine-grained open classes of named entities
  • Survey proposed schemes of fine-grained open classes, for example:
    • 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 2021 or 2022 paper as the final primary source, use the citation chain to find at least two previous papers
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:
• Vladimir Dobrovolskii (2021). Word-Level Coreference Resolution. EMNLP
  • https://arxiv.org/abs/2109.04127
• (Dr. Viktor Hangya, Katharina Hämmerl, Wen Lai)
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* Thursday
  • 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!