Information Extraction
Topics

CIS, LMU München
Winter Semester 2023-2024

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**
  - 6 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!**
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.
• 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

• First presenter:
  • Quickly explain three different workshop series: MUC, ACE, TAC
  • What is a shared task, who is participating, what is done, how is it evaluated
  • Give a brief idea of the types of tasks, ideally with simple examples

• Second presenter: Recognizing Ultra Fine-Grained Entities (RUFES) shared task at TAC 2022

• Third presenter: Streaming Multimedia Knowledge Base Population (SM-KBP) at TAC 2022

• All presenters:
  • Mention best systems
  • Discuss problems solved, motivations and techniques
  • Survey the literature

• 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 recent (ideally 2023) 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
• (Dr. Viktor Hangya, Katharina Hämmerl, Faeze Ghorbanpour)
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!