Seminar Topics: Information Extraction

Matthias Huck, Alexander Fraser

LMU Munich

21 October 2019

Arabic Named Entity Recognition

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Overview:

- Discuss how to build an NER system for Arabic, a member of the Semitic languages family.
- What are the linguistic issues and challenges? How can they be tackled?

Paper:

- Shaalan (2014).
  A Survey of Arabic Named Entity Recognition and Classification.
  Computational Linguistics, 40(2), pp. 469-510.
  `https://doi.org/10.1162/COLI_a_00178`

Clinical Information Extraction with CRFs

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Overview:

- IE for the biomedical domain can be valuable, but is difficult to build, for instance due to a large amount of special terminology.
- How can IE via supervised learning (with CRFs) be performed from free-text clinical reports? How reliable is it and what are domain-specific challenges?

Paper:

- Esuli et al. (2013).
  An enhanced CRFs-based system for information extraction from radiology reports.
  Journal of Biomedical Informatics, 46(3), pp. 425-435.
  `https://doi.org/10.1016/j.jbi.2013.01.006`

Automatic Intent Identification in Email

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Overview:

- Reading and answering emails eats up a considerable proportion of enterprise workers' time.
- Intent classes in enterprise email conversations:
  request information, *schedule meeting*, *promise action*, ...
- Automatic intent identification may increase productivity.
- How to build a state-of-the-art email intent identification system?

Paper:

- Wang et al. (2019).
  Context-Aware Intent Identification in Email Conversations.
  In Proc. of SIGIR, pp. 585-594, Paris, France, July.
  `https://doi.org/10.1145/3331184.3331260`

Structural Segmentation of Email with a Statistical Classifier

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Overview:

- Zones in email body text:
  author, greeting, signoff, reply, forward, signature, advertising, disclaimer, attachment.
- What is needed to automatically segment email messages into zones using a statistical classifier? How well does it work?

Paper:

- Lampert et al. (2009).
  Segmenting Email Message Text into Zones.
  In Proc. of EMNLP, pp. 919–928, Singapore, August.
  `https://aclweb.org/anthology/D09-1096/`

Semantic Role Labeling with RNNs

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Overview:

- SRL: detecting the predicate-argument structure of a sentence.
- Can a recurrent neural network model be adopted to solve SRL as a supervised machine learning task?
- (Please also describe in detail what SRL is and motivate why it may be useful for information extraction.)

Paper:

- Zhou and Xu (2015).
  End-to-end Learning of Semantic Role Labeling Using Recurrent Neural Networks.
  In Proc. of ACL-IJCNLP, pp. 1127-1137, Beijing, China, July.
  `https://www.aclweb.org/anthology/P15-1109/`

Neural Open Information Extraction

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Overview:

- Relation extraction: Obtain triples such as *(Elvis, born_in, Tupelo)* from unstructured text containing that information.
- Open information extraction:
  The IE system is able to discover new types of relations.
- How can neural network models be applied to build an open relation extraction system?

Papers:

- Cui et al. (2018).
  Neural Open Information Extraction.
  In Proc. of ACL (Short Papers), pp. 407-413, Melbourne, Australia, July.
  `https://www.aclweb.org/anthology/P18-2065/`
- Trisedya et al. (2019).
  Neural Relation Extraction for Knowledge Base Enrichment.
  In Proc. of ACL, pp. 229-240, Florence, Italy, July.
  `https://www.aclweb.org/anthology/P19-1023/`

Event Detection for Historical Texts

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Overview:

- Automatic processing of historical text corpora becomes more useful with the current massive digitization of historical sources.
- How should event mentions and types be annotated in historical corpora?
- Which approaches should be chosen to develop a system for automatic event detection and classification?

Paper:

- Sprugnoli and Tonelli (2019).
  Novel Event Detection and Classification for Historical Texts.
  Computational Linguistics, 45(2), pp. 229-265.
  `https://doi.org/10.1162/coli_a_00347`

Hint: Paywalled Literature

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Access to publications behind a paywall can often be provided via the university library.

Try “**E-Medien-Login**”, using your LMU user ID:
`http://www.ub.uni-muenchen.de/ausleihe-online/digitaler-zugriff/e-medien-login/index.html`

Alternatively, search the web for preprint versions.

Questions?

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Thank you for your attention

Matthias Huck

mhuck@cis.lmu.de