Seminar Topics: Information Extraction

Matthias Huck, Alexander Fraser

LMU Munich

25 October 2017

Joint Named Entity Recognition and Disambiguation

Overview:

- NER: detecting text spans of entity mentions and tagging them with coarse-grained types.
- NED: mapping mentions to entities in a knowledge base (KB).
- Can both be done jointly rather than in two separate stages?

Paper:

- **J-NERD: Joint Named Entity Recognition and Disambiguation with Rich Linguistic Features**.
  Dat Ba Nguyen, Martin Theobald, Gerhard Weikum.
  TACL 2016.

  `https://transacl.org/ojs/index.php/tacl/article/view/698`

Recommended prior knowledge:

- Some basic understanding of CRFs.

Named Entity Recognition with Neural Networks

Overview:

- What would a state-of-the-art neural model for NER look like?
- Using a hybrid LSTM-CNN architecture.
- With word- and character-level features.
- And employing publicly available word embeddings.

Paper:

- **Named Entity Recognition with Bidirectional LSTM-CNNs**.
  Jason P.C. Chiu, Eric Nichols.
  TACL 2016.
  `https://transacl.org/ojs/index.php/tacl/article/view/792`

Recommended prior knowledge:

- Some basic understanding of RNNs & word embeddings.

Relation Classification with Neural Networks

Overview:

- Relation classification: identifying the semantic relation between two entities in text.
- What would a state-of-the-art neural model for relation classification look like?

Paper:

- **Relation Classification via Multi-Level Attention CNNs**. Linlin Wang, Zhu Cao, Gerard de Melo, Zhiyuan Liu. ACL 2016.

  `http://www.aclweb.org/anthology/P/P16/P16-1123.pdf`

Recommended prior knowledge:

- Some basic understanding of CNNs.
- Attention mechanism as in sequence-to-sequence learning.

Relation Extraction via Reading Comprehension

Overview:

- Relation extraction systems can be used to populate knowledge bases with facts from an unstructured text corpus.
- Challenging when the types of facts (relations) are not predefined.
- How can we generalize to unseen relations?
  (Zero-shot learning problem.)
- Can relation extraction be reduced to reading comprehension?

Paper:

- **Zero-Shot Relation Extraction via Reading Comprehension**.
  Omer Levy, Minjoon Seo, Eunsol Choi, Luke Zettlemoyer.
  CoNLL 2017.

  `http://www.aclweb.org/anthology/K/K17/K17-1034.pdf`

(ADVANCED TOPIC.)

Open-Domain Question Answering

Overview:

- The answer to any factoid question is a text span in Wikipedia.
- Document retrieval: finding the relevant articles.
- Machine comprehension: identifying the answer spans.
- *Machine reading at scale*:
  How to build a modern large-scale QA system?

Paper:

- **Reading Wikipedia to Answer Open-Domain Questions**.
  Danqi Chen, Adam Fisch, Jason Weston, Antoine Bordes.
  ACL 2017.
  `http://www.aclweb.org/anthology/P/P17/P17-1171.pdf`

Dialogue Agents for Information Access

Overview:

- How to build a modern dialogue agent which helps users search knowledge bases without composing complicated queries?

Paper:

- **Towards End-to-End Reinforcement Learning of Dialogue Agents for Information Access**.
  Bhuwan Dhingra, Lihong Li, Xiujun Li, Jianfeng Gao, Yun-Nung Chen, Faisal Ahmed, Li Deng.
  ACL 2017.
  `http://www.aclweb.org/anthology/P/P17/P17-1045.pdf`

(ADVANCED++ TOPIC.) Recommended prior knowledge:

- Calculus and probability theory.
- RNNs.
- Reinforcement learning.

Extracting Structured Information from Conversations

Overview:

- Given a dialogue between a customer and a waiter in a restaurant, how could a computer understand the order?
- Sequence-to-sequence learning task: “translating” from the natural language conversation to a structured data record.

Paper:

- **May I take your order? A Neural Model for Extracting Structured Information from Conversations.**
  Baolin Peng, Michael Seltzer, Y.C. Ju, Geoffrey Zweig, Kam-Fai Wong.
  EACL 2017.
  `http://www.aclweb.org/anthology/E/E17/E17-1043.pdf`

Recommended prior knowledge:

- Sequence-to-sequence learning (encoder-decoder approach).

Automatic Biomedical Knowledge Extraction

Overview:

- How to automatically discover important facts by mining biomedical literature?
- Named entity extraction, relation extraction, and ranking of extracted insights in the biomedical domain.

Paper:

- **An Insight Extraction System on BioMedical Literature with Deep Neural Networks**.
  Hua He, Kris Ganjam, Navendu Jain, Jessica Lundin, Ryen White, Jimmy Lin.
  EMNLP 2017.
  `http://www.aclweb.org/anthology/D/D17/D17-1284.pdf`

Credibility Prediction of User Statements

Overview:

- Discussions in online communities are often plagued by inaccuracies and misinformation.
- Can the credibility of drug side-effect statements in health communities be assessed automatically?

Paper:

- **People on Drugs: Credibility of User Statements in Health Communities.**
  Subhabrata Mukherjee, Gerhard Weikum, Cristian Danescu-Niculescu-Mizil.
  KDD 2014.
  `https://dl.acm.org/citation.cfm?id=2623714`

Event Detection in Social Media

Overview:

- Information on many real-world events appears in social media before any traditional news agencies report.
- Can disruptive events be automatically detected in the streamed social media data?

Paper:

- **Can We Predict a Riot? Disruptive Event Detection Using Twitter**.
  Nasser Alsaedi, Pete Burnap, Omer Rana.
  ACM Transactions on Internet Technology 2017.
  `https://dl.acm.org/citation.cfm?id=2996183`

A Web-scale Probabilistic Knowledge Base

Overview:

- Exploring automatic methods for constructing knowledge bases.
- How to extract facts from the Web and combine them with existing prior knowledge?
- Google's *Knowledge Vault*: about 38 times bigger than other automatically constructed KBs (in 2014).

Paper:

- **Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion**.
  Xin Luna Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, Wei Zhang.
  KDD 2014.
  `https://dl.acm.org/citation.cfm?id=2623623`

Hint: Paywalled Literature

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?

Thank you for your attention

Matthias Huck

mhuck@cis.lmu.de