Seminar Topics: Information Extraction

English topics!

Alexandra Chronopoulou

achron@cis.lmu.de

Topic: Offensive Language Detection

- Offensive language is very common on social media platforms. It has various forms, such as **hate speech** (targeted to a group), **cyberbullying** (targeted to an individual), **aggression**.
- Target: Automatic identification of offensive language
- The task is usually formulated as a supervised classification problem
- Datasets are created from posts annotated with respect to the presence of some form of abusive content
- This topic covers a shared task: SemEval 2019 Task 6

Topic: Offensive Language Detection

1. Overview:
   - Zampieri et al., 2019, **SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval)** In *Proceedings of the International Workshop on Semantic Evaluation*
2. Rule-based approach & deep-learning approach
   - Han et al., 2019, **jhan014 at SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media** In *Proceedings of the International Workshop on Semantic Evaluation*
   - Zhang et al., 2019, **MIDAS at SemEval-2019 Task 6: Identifying Offensive Posts and Targeted Offense from Twitter** In *Proceedings of the International Workshop on Semantic Evaluation*
3. State-of-the-art deep learning (BERT) approach
   - Liu et al., 2019, **NULI at SemEval-2019 Task 6: Transfer Learning for Offensive Language Detection using Bidirectional Transformers** In *Proceedings of the International Workshop on Semantic Evaluation*

Open-Domain Question-Answering

- Knowledge bases (KB) access annotated relational data by enabling queries such as (DANTE, born-in, **X**)

Open-Domain Question-Answering

- Knowledge bases (KB) access annotated relational data by enabling queries such as (DANTE, born-in, **X**)
- In practice though, we often need to **extract** relational data from text to populate these knowledge bases

Open-Domain Question-Answering

- Knowledge bases (KB) access annotated relational data by enabling queries such as (DANTE, born-in, **X**)
- In practice though, we often need to **extract** relational data from text to populate these knowledge bases
- So, we need entity extraction, entity linking, relation extraction - but these methods need **supervised** data and **fixed schemas**

Open-Domain Question-Answering

- Knowledge bases (KB) access annotated relational data by enabling queries such as (DANTE, born-in, **X**)
- In practice though, we often need to **extract** relational data from text to populate these knowledge bases
- So, we need entity extraction, entity linking, relation extraction - but these methods need **supervised** data and **fixed schemas**
- Pretrained language models have become increasingly important for NLP. They are optimized to predict a *masked word* anywhere in a sentence and appear to store *vast amounts* of linguistic knowledge

Open-Domain Question-Answering

- Knowledge bases (KB) access annotated relational data by enabling queries such as (DANTE, born-in, **X**)
- In practice though, we often need to **extract** relational data from text to populate these knowledge bases
- So, we need entity extraction, entity linking, relation extraction - but these methods need **supervised** data and **fixed schemas**
- Pretrained language models have become increasingly important for NLP. They are optimized to predict a *masked word* anywhere in a sentence and appear to store *vast amounts* of linguistic knowledge
- We can now query **pretrained language models** for relational data:

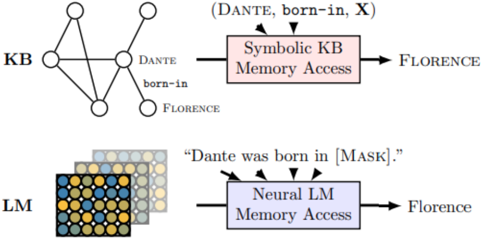

Fig. from Petroni et al., 2019

Open-Domain Question-Answering

1. Neural vs count-based distrib. methods on lexical semantics tasks
   - Baroni et al., 2014, **Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors** In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*
2. Baselines
   - Relation extraction model: Sorokin and Gurevych, 2017, **Context-Aware Representations for Knowledge Base Relation Extraction** In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*
   - Open-Domain QA model: Chen et al., 2017, **Reading Wikipedia to Answer Open-Domain Questions** In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*
3. State-of-the-art deep-learning model
   - Petroni et al., 2019, **Language Models as Knowledge Bases?** In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*

Nested Named Entity Recognition

- **Named entity recognition** is the task of *identifying text spans associated with proper names* and classifying them according to their semantic class such as *person*, *organization*, etc
- **Mention detection**: text spans *referring to named, nominal or prominal entities* are identified and classified according to their semantic class
- Most methods suffer from an inability to handle *nested* named entities, *nested* entity mentions, or both
- In the Fig. below, a PERSON named entity is nested in an entity mention of type LOCATION

… [the burial site of [Sheikh Abbad]$_{\text{PERSON}}$]$_{\text{LOCATION}}$ is located …

Fig. from Katiyar and Cardie, 2018.

- Most existing methods would **miss the nested entity** - and nested entities are fairly **common**

Nested Named Entity Recognition

1. Mention hypergraph model for nested entity detection
   - Lu and Roth, 2015, **Joint Mention Extraction and Classification with Mention Hypergraphs** In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*
2. Neural network-based methods for *simple* NER
   - Chiu and Nichols, 2016, **Named Entity Recognition with Bidirectional LSTM-CNNs** In *Transactions of the Association for Computational Linguistics*
   - Lample et al., 2016, **Neural Architectures for Named Entity Recognition** In *Proceedings of the North American Chapter of the Association for Computational Linguistics*
3. Neural-network based approach for *nested* NER
   - Katiyar and Cardie, 2018, **Nested Named Entity Recognition Revisited** In *Proceedings of the North American Chapter of the Association for Computational Linguistics*

Coreference Resolution

- Coreference resolution is an important task for NLP

Coreference Resolution

- Coreference resolution is an important task for NLP
- But what exactly is this task?

Coreference Resolution

- Let's have a look at Bob who is talking with his AI friend Alice:

Coreference Resolution

- Let's have a look at Bob who is talking with his AI friend Alice:

- There are several implicit references in the last message from Bob: "**she**" refers to the same entity as "My sister": Bob's sister

Coreference Resolution

- Let's have a look at Bob who is talking with his AI friend Alice:

- There are several implicit references in the last message from Bob: "**she**" refers to the same entity as "My sister": Bob's sister
- "**he**" refers to the same entity as "a friend called John": Bob's sister's friend

Coreference Resolution

- Let's have a look at Bob who is talking with his AI friend Alice:

- There are several implicit references in the last message from Bob: "**she**" refers to the same entity as "My sister": Bob's sister
- "**he**" refers to the same entity as "a friend called John": Bob's sister's friend
- The process of linking together mentions that relates to real world entities is called *coreference resolution*

Coreference Resolution

1. Mention-pair classifier
   - Clark and Manning, 2016, **Improving Coreference Resolution by Learning Entity-Level Distributed Representations** In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*
2. Latent-tree and mention ranking models
   - Martschat and Strube, 2015, **Latent Structures for Coreference Resolution** In *Transactions of the the Association for Computational Linguistics*
   - Durrett and Klein, 2013, **Easy Victories and Uphill Battles in Coreference Resolution** In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*
3. Deep learning method
   - Lee et al., 2017, **End-to-end Neural Coreference Resolution** In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*