Seminar Topics: Information Extraction

English topics!

Alexandra Chronopoulou

achron@cis.lmu.de
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:

2. Rule-based approach & deep-learning approach

3. State-of-the-art deep learning (BERT) approach
Knowledge bases (KB) access annotated relational data by enabling queries such as (DANTE, born-in, X)
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.
Knowledge bases (KB) access annotated relational data by enabling queries such as \((\text{DANTE}, \text{born-in}, \text{X})\).

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 \((\text{DANTE}, \text{born-in}, X)\)
- 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 \( (\text{DANTE}, \text{born-in}, X) \)
- 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

3. State-of-the-art deep-learning model
**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]_PERSON]_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:

[Chat messages]
- Bob: My sister has a friend called John
- Alice: She thinks he is so funny
- Bob: Really, tell me more about him

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

- Let's have a look at Bob who is talking with his AI friend Alice:
  
  ![](image1.png)

- 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
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

3. Deep learning method