

Annotated datasets for NER

TOPIC: Training data for Named Entity Recognition

- Give a brief overview of available annotated datasets for NER
 - i.e. the data we need to train models with full supervision
- Do you think this is enough data to train good supervised models?
 - give us some results that support your answer
 - what about using unsupervised learning?

Sources and possible papers:

- *Nadeau and Sekin, A survey of named entity recognition and classification* <http://nlp.cs.nyu.edu/sekine/papers/li07.pdf>

Annotated data for Medical NER

TOPIC: Named Entities in the CLEF-eHEALTH challenge

- Give an overview of the CLEF-eHEALTH challenge
- Talk about NER in this Challenge (Task 1)
- Present the training data provided for medical NER
 - Which set of classes are annotated?
- How can you use this data to train a classifier (e.g. a linear model)?

Sources and possible papers:

- <https://sites.google.com/site/clefehealth2016/>
- <https://sites.google.com/site/clefehealth2015/>
- ...

Supervised NER

TOPIC: CNN for Named Entity Recognition

- Explain feedforward neural networks (FFNN).
- Explain convolutional neural networks. In which respect are they different from FFNN.
- Discuss applications of these two architectures for NER. Which one works better and why?
 - Sources and possible papers:
 - Collobert et al. (2011), *Natural Language Processing (Almost) from Scratch*, *Journal of Machine Learning Research*, p. 2461-2505.

Supervised NER

TOPIC: LSTM for Named Entity Recognition

- Explain recurrent neural networks (RNN).
- Explain long-short term memory networks (LSTM). Why are LSTM useful?
- Discuss applications of LSTM for NER. Why is it a good idea to add a conditional random field (CRF) on top of the LSTM?
 - Sources and possible papers:
 - *Lample et al. (2016), Neural Architectures for Named Entity Recognition, Proc. ACL.*

Bilingual Sentence Extraction

TOPIC: Description of the BUCC shared task

- What is the BUCC shared task and which NLP problem does it address?
 - How is the training and test data built for that task?
 - What impact has topic selection on the difficulty of the task?
 - Give an overview of submitted systems and results.
- Sources and possible papers:
 - *Zweigenbaum et al. (2017), Overview of the Second BUCC Shared Task: Spotting Parallel Sentences in Comparable Corpora, Proc. Workshop on Building and Using Comparable Corpora.*

Bilingual Sentence Extraction

TOPIC: A neural network architecture for bilingual sentence extraction.

- How can neural network architectures be used for bilingual sentence extraction?
- What are bilingual word embeddings and how are they used for candidate filtering?
- How does the neural classifier work?
 - Sources and possible papers:
 - *Gregoire et al. (2017), BUCC 2017 Shared Task: a First Attempt Toward a Deep Learning Framework for Identifying Parallel Sentences in Comparable Corpora, Proc. Workshop on Building and Using Comparable Corpora.*

Supervised NER

TOPIC: Weakly Supervised Named Entity Recognition

- Starting from a few examples ("seed examples"), how do you automatically build a named entity classifier?
- This is sometimes referred to as "bootstrapping"
- What are the problems with this approach?
- How do you block the process from generalizing too much?
- Should we use weak supervision instead of (full) supervision for NER
- give us some results that support your answer
 - Sources and possible papers:
 - *Nadeau and Sekin, A survey of named entity recognition and classification* <http://nlp.cs.nyu.edu/sekine/papers/li07.pdf>

Crowd-sourcing for NER

TOPIC: Crowd-sourcing with Amazon Mechanical (AMT)

- AMT's motto: artificial artificial intelligence
- Using human annotators to get quick (but low quality) annotations
- What are the pros and cons of this approach?
- Present how NER data is collected using AMT
- How well do NER systems perform when trained on this data?

Sources and possible papers:

- Finin et al., *Annotating named entites in Twitter data with crowdsourcing.*
http://ebiquity.umbc.edu/_file_directory_/papers/483.pdf

NER Domain Adaptation

TOPIC: Domain adaptation and failure to adapt

- What is the problem of domain adaptation?
- How is it addressed in statistical classification approaches to NER?
- How well does it work

Sources and possible papers:

- Daume III, *Frustratingly Easy Domain Adaptation*
<http://www.umiacs.umd.edu/~hal/docs/daume07easyadapt.pdf>

Classification-based Citation Parsing

TOPIC: parsing citations using classifiers

- How is the citation parsing problem formulated using classifiers ?
- What sort of information is available?
- What does the training data look like?
- What sorts of downstream applications are based on citation parsing ?

Sources and possible papers:

- Peng et al., *Information extraction from research papers using conditional random fields* <http://dl.acm.org/citation.cfm?id=1142104>