Topic Presentation: Essence of a claim

Le Anh Phuong - 2019.15.04

What is a claim?

- A statement essentially arguable, but used as a primary point to support or prove an argument (https://literarydevices.net/claim/)
- Central part of all arguments
- Usually needs some support to make a full argument - premises, evidence or justifications

The Task

- Given a sentence, classify whether or not it contains a claim
  - If one or more tokens within the sentence were labeled as claim
  - Keep content of the entire document to be able to retrieve information about the context of (non-)claims
- Compare the influence of different type of information (lexical - word/vocabulary, syntactical - syntax and others) across datasets

6 English Datasets (Corpora)

- Various genres (VG)
- Web discourse (WD)
- Persuasive essays (PE)
- Online comments (OC)
- Wiki talk pages (WTP)
- Micro texts (MT)

Machine Learning Algorithm

- Logistic Regression + one or more features
  - Structure (position, length, punctuation)
  - Lexical (lower-cased unigrams)
  - Syntax (grammatical)
  - Discourse (debate - encoded information extracted by discourse parser)
  - Embedding (summation of word embedding)
- Deep Learning Approaches
  - Convolutional Neural Net of Kim
  - Pre-trained word embeddings CNN:w2vec / CNN:rand

Results: In-domain experiments

- Features comparison
  - Lexical, embedding and syntax features are helpful
  - Structural features did not help
  - Discourse features only contribute significantly on MT
- LR (logistic regression) + syntax feature and CNN:rand perform virtually identical
- Dataset comparison:
  - Could not search for correlation btw performance because of different nature of inter-annotator
  - PE and MT has better results and good inter-annotator agreement

Results: Cross-domain experiments

- Biggest performance drops on the datasets which performed best in the in-domain setting (MT & PE)
- Lowest scoring datasets: OC & WTP - the differences are small when trained on suitable dataset
- Best of feature based approach outperforms best of deep learning approach
- Training on VG or OC seems the best when unknown domain of test data while MT gives best results as target domain
- Mixed sources works better than single source (larger dataset)