Classify Conclusion / Premise

20.05.2019 - Annie & Lukas

Convert json to XMl

- # conclusions: 11556 (2 duplicates - 0.017%), # premises: 35871 (928 duplicates - 2.5%)
- Only takes distinct sentence

Splitter

statistic

	Topic	Unit total	Conclusion	Premise
All	460	46497	11554 (25%)	34943 (75%)
Train	367 (80%)	36761 (79%)	9122 (25%)	27639 (75%)
Test	93 (20%)	9736 (21%)	2432 (25%)	7304 (75%)

Train classifier (all)

- Baseline: 75%
- token pos oversampling: 91%
- token pos undersampling: 90%
- token oversampling: 87%
- token undersampling: 86%
- pos oversampling: 90%
- pos undersampling: 88%

screenshots of Weka - tokenpos, token, pos arff

Should we use context?

- In debatepedia, the context is not given (conclusion and premises are separated)
- In parliamentary debate, the conclusion maybe at the beginning or the end of the speech => we may use another training set to take advantage of the context

Should we use topic?

- On average, each topic is related to 100 sentences, 25% is conclusion and the topic is concrete & defined. It is not wise to train topic one by one
- But it maybe an idea to train on the whole dataset (put 80% of every topic in training, 20% on testing), and include topic as a feature and generate corresponding value in feature vectors.
- Topic matching between 2 datasets (debatepedia & parliamentary debate) is also considered

Some idea for topic matching

Topics from debatepedia

- Exact matching
- Weki topic matching
- Synonym matching