NJUST @ CLSciSumm-18

Shutian Ma, Heng Zhang, Jin Xu, Chengzhi Zhang

Department of Information Management, Nanjing University of Science and Technology, Nanjing, China, 210094

3rd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2018)
Teammate
Task 1A

- Select suitable features for classifiers
- Integrate results via weighted voting system
Task 1A

- Citation Text Preprocess
- Feature Selection
- Parameter Setting
- New Classifier
Task 1A

- Citation Text Preprocess
- Feature Selection
- Parameter Setting
- New Classifier

**Paper [1]**

Like others, we have assumed lexical semantic classes of verbs as defined in \textit{Levin (1993)} (hereafter Levin), which have served as a gold standard in computational linguistics research (Dorr and Jones, 1996; Kipper et al., 2000; Merlo and Stevenson, 2001; Schulte im Walde and Brew, 2002). Levin’s classes form a hierarchy of verb groupings.

**Paper [2]**

The system described in this paper is similar to the MENE system of (Borthwick, 1999). \textit{It} uses a maximum entropy framework and classifies each word given its features.
Task 1A

- Citation Text Preprocess
- Feature Selection
- Parameter Setting
- New Classifier

Iteratively evaluate a candidate subset of selected features set + fixed feature set
Task 1A

- Citation Text Preprocess
- Feature Selection
- Parameter Setting
- New Classifier

Different Voting Weights of Precision, Recall and F1 Oriented for 4-Classifiers and 3-Classifiers System
Task 1A

- Citation Text Preprocess
- Feature Selection
- Parameter Setting
- New Classifier: XGBoost, gradient boosted decision trees
Task 1B

- Manual Dictionary
- LLDA
- XGBOOST
- Manual dictionary + LLDA
- POS dictionary + LLDA
Task 1B

Dictionaries of five facets

- Manual Dictionary
- LLDA
- XGBOOST
- Manual dictionary + LLDA
- POS dictionary + LLDA
Task 1B

- Manual Dictionary
- LLDA
- XGBOOST
- Manual dictionary + LLDA
- POS dictionary + LLDA

✓ Supervised topic model
✓ We assume that each identified facet is a topic label.
✓ We firstly train LLDA on the training data and the dimension number is five. Then, we apply this trained model to do predictions over testing data.
Task 1B

- Manual Dictionary
- LLDA
- XGBOOST
- Manual dictionary + LLDA
- POS dictionary + LLDA
Task 1B

- Manual Dictionary
- LLDA
- XGBOOST
- Manual dictionary + LLDA
- POS dictionary + LLDA

We use the dictionary-labeled testing data for LLDA prediction.
Task 2

- Group sentences into different clusters
- Use features to score sentence from each cluster and combine them
We find that when people write summaries like abstract, they will start with some fixed phrases, such as “this paper”, “in this paper” or “we”. If the first sentence doesn’t have these fixed phrases, it will be about motivation of this paper for most of the time. Meanwhile, the last sentence are usually about results or conclusions.
Task 2

- Split abstract sentences into groups
- Each identified text span is selected into different groups based on their similarity with the grouped abstract sentences, linear sum of Jaccard, IDF and TFIDF similarities
- Group sentence based on score and combine them into summary

\[ Score_i = 2.5S_{\text{Jaccard}} + 2.5S_{\text{IDF}} + 2.5S_{\text{TFIDF}} + 1.25S_{\text{Length}} + 1.25S_{\text{Position}} \]
Thanks!

Shutian Ma: mashutian0608@hotmail.com
Heng Zhang: 525696532@qq.com
Jin Xu: xujin@njust.edu.cn
Chengzhi Zhang: zhangcz@njust.edu.cn