Engaging Content
Engaging People

Using WordNet for Query Expansion: ADAPT @ FIRE 2016 Microblog Track

Wei Li , Debasis Ganguly, Gareth J.F. Jones
ADAPT Centre, Dublin City University, Ireland

Ireland's European Structural and Investment Funds Programmes 2014-2020
Co-funded by the Irish Government and the European Union
European Union
European Regional Development Fund
SFI Science Foundation Ireland

The ADAPT Centre is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

Outline

- Task Summary
- Experimental Methods
- Results
- Conclusions and Further Work

Task Summary

- Identify relevant tweets posted during a recent disaster event for a set of topics seeking certain types of information.
- Identify relevant tweets with high precision as well as high recall.

Method

Challenges:

- query-document mismatch problems arising from short length of tweets
- differing use of vocabulary in the topics and the tweets

Our Proposal:

- query expansion based on WordNet

WordNet:

- an electronic lexical database: synonyms, hypernyms or hyponyms
- long regarded as a potentially useful resource for query expansion in information retrieval

Method

Data gathering:

- Downloaded 49,894 of 50,068 listed tweet ids

Indexing:

Tweets indexed for search using Lucene:

- entries from a list of 655 stop words removed;
- Porter stemmer applied to all words;
- BM25 model used for retrieval with k1=1.2, b=0.75.

Method

Two experiments conducted based on WordNet:

- Automatic method
- Semi-automatic method

- For both methods, synonyms for each topic term limited to a maximum of 20.
  - some terms received less synonyms

Experiment One

Automatic method:

- remove stop words from each topic
- use WordNet to generate the synonyms for each item in every topic
- use synonyms to expand the query terms
- apply expanded topic to Lucene system to search with BM25

Note: The original search topics is made up of the combination of title and narrative fields of each topic.

Experiment Two

Semi-automatic method:

- Use the original topic to search and obtain a ranked list
- Go through top 30 tweets, select 1-2 relevant tweets to perform query expansion.
- Remove stop words and duplicate terms from the selected tweets, add the remaining terms to the original topic
- Applied WordNet again on the expanded topics and find synonyms for these terms
- Add synonyms to expanded topic to generate new topic
- Search again

Results

- Our automatic run received the third place among submission, however with the best MAP value
- Our semi-automatic run obtained the overall first place

Run Type	Run Name	Rank	P@20	R@100	MAP@100	MAP
Auto Run	iiest_saptarashmi_bandyopadhyay_1	1	0.4357	0.3420	0.0869	0.1125
Auto Run	dcu_fmt16_1	3	0.3786	0.3578	0.1103	0.1103
Semi-auto Run	dcu_fmt16_2	1	0.4286	0.3445	0.0815	0.0815
Semi-auto Run	iitbhu_fmt16_1	2	0.3214	0.2581	0.0670	0.0827

Conclusions and Further Work

Conclusions

- Use of WordNet as an external resource for query expansion showed positive results for this task.
- Augments the original query to include symonym words which are more effective at matching relevant tweets.

Further Work

- Use document expansion to expand tweets based on external resources.
- Use WordNet to identify hypernyms or hyponyms for each topic term as additional expansion items.