Improved Automatic Keyword Extraction Based on TextRank using Domain Knowledge

GUANGYI LI, HOUFENG WANG

INSTITUTE OF COMPUTATIONAL LINGUISTICS

PEKING UNIVERSITY

Outline

- Introduction
- General Framework
- Improvement by Domain Knowledge
- Experiments
- Conclusion

Introduction

- Keywords (Keyphrases) consist of one word or several words
- Keywords summarize topics and ideas of an article
- Keywords can benefit many NLP applications, such as text categorization, document clustering, search engine, etc.
- SemEval 2010 Shared Task 5: Keyword Extraction for scientific articles

Related Work

- Candidate Selection
  - N-gram
  - Part-of-Speech
  - NP-Chunk
- Choosing from Candidates
  - Statistical methods: tf-idf, PMI, etc.
  - Supervised methods: ME, NB, SVM, CRF
  - Unsupervised method: TextRank

General Framework

- Candidate Selection By Document Frequency Accessor Variety
- Ranking candidates By phrase-based TextRank
- Improvement with Domain Knowledge

Candidate Selection

- Rule-based candidate selection performs worse for Chinese
- Accessor Variety (AV) shows how often a phrase appears as a whole
- Accessor Variety is the number of different words appear left or right to the phrase
- Accessor Variety doesn't work well for low-frequency phrases

Candidate Selection

- We find that keywords are usually surrounded by common words. Common words have high document frequency
- Therefore, we propose Document Frequency Accessor Variety

$$DFAV_L = \sum_{w \in S_L} \log doc\_freq(w)$$

$$DFAV_R = \sum_{w \in S_R} \log doc\_freq(w)$$

$$Score(phr) = DFAV_L(phr) \times DFAV_R(phr)$$

Phrase-based TextRank

- TextRank is inspired by PageRank. A word is a node, and co-occurrence between words is an edge
- Previous work ranks words and use top-ranked words to generate keywords
- Not every word in keywords can rank top
- So we involve candidate phrases into the graph

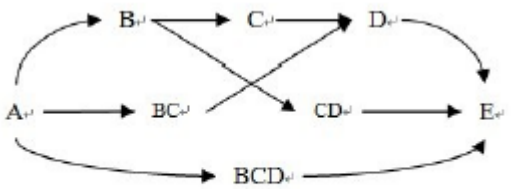

Fig. 1. Neighboring Graph

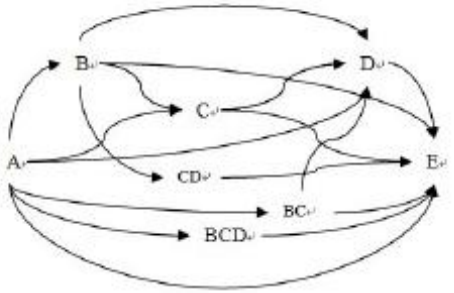

Fig. 2. Phrase Network Graph

Improvement by Domain Knowledge

- We obtained a large number of abstracts and keywords of scientific articles from cnki.net
- We show how to improve keyword extraction of a certain domain with domain knowledge
- Length of Keywords
- Components of Keywords
- High-frequency Keywords

Length of Keywords

Table 1. Average Length of Keyword in Different Domains

Domain	ethnology	petroleum	mathematics	international law
Ave. Len.	3.54	4.16	4.62	4.48

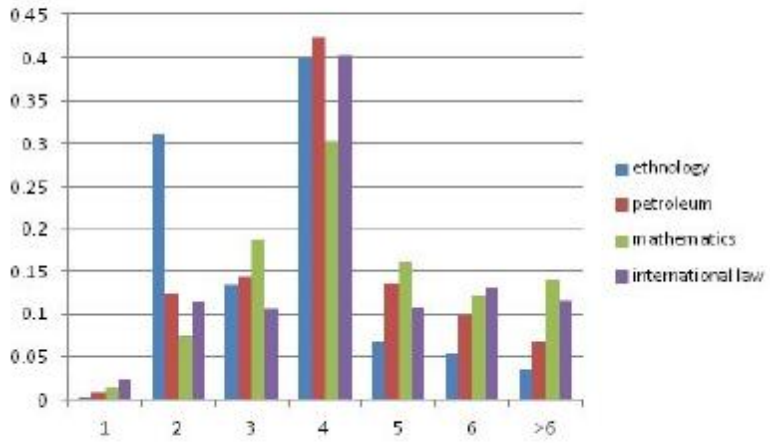

Fig. 3. Distribution of Length of Keyword

- Edge Weight for TextRank
- Multiplied factor to score

Components of Keywords

- Some words are unique to a domain
  - 随机(random) is unique to mathematics domain
  - 文化(culture) is unique to ethonology domain
- Some words only appear in specific position of keywords
  - 系统(system) only appear at the end of keywords
- Extract such rules for candidates selection

High-frequency Keywords

- High-frequency Keywords are a natural thesaurus for the domain
- Increase TextRank score of high-frequency keyword by the cube root of its frequency

Experiments

- Titles, abstracts, keywords from cnki.net
  - 100 articles for each domain as test set
  - Keywords of 1000 articles for each domain as domain knowledge
- Segmentation and part-of-speech tagging by a perceptron-based tool
- Evaluation
  - P, R, F1 of top 5, 10, 15

Experimental Results

	ethnology			petroleum			mathematics			international law		
	Top5	Top10	Top15	Top5	Top10	Top15	Top5	Top10	Top15	Top5	Top10	Top15
TF-IDF	0.243	0.234	0.201	0.108	0.141	0.148	0.115	0.122	0.127	0.211	0.189	0.166
TextRank	0.312	0.249	0.201	0.179	0.184	0.173	0.167	0.176	0.173	0.287	0.238	0.197
+component	0.319	0.253	0.199	0.176	0.184	0.176	0.170	0.176	0.176	0.285	0.239	0.196
+length	0.326	0.256	0.203	0.181	0.186	0.176	0.172	0.179	0.178	0.290	0.242	0.197
+high-freq	0.342	0.258	0.205	0.202	0.201	0.180	0.180	0.187	0.183	0.300	0.249	0.199

Conclusion

- This paper
  - An effective way to extract keywords for Chinese scientific articles
  - Domain knowledge can improve keyword extraction
- Future Work
  - Improve precision of candidate selection
  - Exploit more domain characteristics

THANK YOU!