Keyphrase extraction. Abstracts instead of full papers

Svetlana Popova

Saint-Petersburg State University and ITMO University, Russia

Vera Danilova

Autonomous University of Barcelona, Spain

What is KeyPhrase?

- A word or a group of words, which reflects the domain-specific of the text

Why is KeyPhrase?

- data indexing, clustering and classification of documents, meta-information extraction, automatic ontologies creation etc.

Approaches

- Most frequent sequencys
- Single word ranking, best words selection and concatenation of best words following each other in the text
- Two stages: a selection stage, when candidate phrases are selected, and a classifying or ranking stage.

Candidate selection

- n-grams, nouns and adjectives (PoS)
- position at the beginning of a document (works for academic papers)
- too many candidates negatively influence ranking

  (W. You, D. Fontaine and J.-P. Barhes, “An automatic keyphrase extraction system for scientific documents,” In: Knowl Inf Syst 34, 2013, pp. 691-724)

- the size of obtained sequences is limited to 4-5

Main idea

using of **abstracts instead of full texts** allows to improve the results obtained by processing full texts or abstracts with introduction and conclusion section

DataSet and Evaluation

- automatic keyphrase extraction task at the Workshop on Semantic Evaluation 2010 (SemEval-2010)
- 244 scientific articles with keyphrase annotations made by authors and readers (we use only readers)
- categories: C2.4 (Distributed Systems), H3.3 (Information Search and
- Retrieval), I2.11 (Distributed Artificial Intelligence—Multiagent Systems) and J4 (Social and Behavioral Sciences—Economics).

DataSet	Total Docs.	Topic			
		C	*H*	*I*	*J*
TRAIN	144	34	39	35	36
TEST	100	25	25	25	25

Text 1	Distributed Task Allocation in Social Networks This paper proposes a new variant of the task allocation problem, where the agents are connected in a social network and tasks arrive at the agents distributed over the network. We show that the complexity of this problem remains NPhard. Moreover, it is not approximable within some factor. We develop an algorithm based on the contract-net protocol. Our algorithm is completely distributed, and it assumes that agents have only local knowledge about tasks and resources. We conduct a set of experiments to evaluate the performance and scalability of the proposed algorithm in terms of solution quality and computation time. Three different types of networks, namely small-world, random and scale-free networks, are used to represent various social relationships among agents in realistic applications. The results demonstrate that our algorithm works well and that it scales well to large-scale applications.
Text 1. Keyphrases labeled by the readers (gold standard)	[social network, multiag system, behavior, strateg agent, social relationship, interact, task alloc, commun messag, util, algorithm, alloc]

Evaluation

$$\text{Precision} = (C \cap G)/G,\ \text{Recall} = (C \cap G)/C,$$

$$\text{F-score} = (2*\text{Precision}*\text{Recall})/(\text{Precision}+\text{Recall}),$$

BEST ALGORITHM PERFORMANCE ON SEMEVAL COLLECTION

Extracted keyphrases (number)	Algorithm	Precision	Recall	F-score
Top-15 keyphrases	HUMB	0.212	0.264	0.235
Top-10 keyphrases	HUMB	0.248	0.206	0.225
Top-5 keyphrases	HUMB	0.304	0.126	0.178

Candidate keyphrase extraction

- Only abstracts were used (“ABSTRACT” indicated the start point and such phrases as “Categories and Subject Descriptors”, “Organization and Design” or “INTRODUCTION” defined the end point.)
- Stanford POS tagger tool, Porter stemmer
- Phrases could not contain punctuation marks (:;,./<>?!@#$%^&*()=+\|") and stop words
- extracted 52 most frequent patterns for the gold standard from the TRAIN: (NN_NN, NN, JJ_NN, NN_NN_NN, JJ_NN_NN, VBN_NN, NN_VBN_NN, VBG_NN, VBN_NN_NN, NN_VBG, NN_JJ_NN, NN_IN_NN_NN_NN, NN_NN_NN_NN, NNS_NN_NN, JJ_NN_NN_NN, NN_VBG_NN, JJ_JJ_NN,….

Results without ranking

- EVALUATION OF KEY PHRASE EXTRACTION USING PATTERNS BUILT ON THE BASIS OF THE ¨TRAIN¨ GOLD STANDARD

DataSet	Precision	Recall	F-score
TEST	0.11	0.67	0.19

Results with ranking (step 1)

- EVALUATION OF KEYPHRASE EXTRACTION. SINGLE-WORD PATTERNS ARE EXCLUDED. IF ONE PHRASE IS A SUBPHRASE OF ANOTHER, WE LEAVE ONLY THE ONE WITH THE HIGHEST ABSOLUTE FREQUENCY VALUE

DataSet	Precision	Recall	F-score
TEST	0.24	0.41	0.30

- EVALUATION OF KEYPHRASE EXTRACTION. SINGLE-WORD PATTERNS ARE EXCLUDED.

DataSet	Precision	Recall	F-score
TEST	0.12	0.52	0.19

Results with ranking (step 1, 2)

- EVALUATION OF KEYPHRASE EXTRACTION USING PATTERNS OF LENGTH 2. Ranking step 1.

DataSet	Precision	Recall	F-score
TEST	0.24	0.41	0.30

- EVALUATION OF KEYPHRASE EXTRACTION USING PATTERNS OF LENGTH 2. Ranking step 2.

Extracted keyphrases (number)	Precision	Recall	F-score
Top-15 keyphrases	0.25	0.30	0.27
Top-10 keyphrases	0.28	0.23	0.26
Top-5 keyphrases	0.34	0.14	0.20

Stop Words

- EVALUATION OF EXTRACTION QUALITY AFTER RANKING. KEYPHRASES OF LENGTH 1 WERE DISCARDED. EXTENDED LIST OF STOP WORDS WAS USED

Extracted keyphrases (number)	Precision	Recall	F-score
No ranking	0.38	0.39	0.38
Top-15 keyphrases	0.34	0.28	0.31
Top-10 keyphrases	0.32	0.23	0.27
Top-5 keyphrases	0.40	0.16	0.23

CONCLUSION

- high capabilities of using abstracts instead of full papers in the task of keyphrase extraction
- abstracts use allows to reduce the number of the selected candidates
- pattern-based keyphrase retrieval gives a Recall of 0.67
- number of pattern-extracted candidates can be reduced by applying an extended stop words

Problems and remarks

- evaluation, collections and gold standards
- abstracts instead full texts but not patterns

Thank you!