Why Keyphrase Extraction?

- Large and growing amounts of research articles indexed by digital libraries.
- Navigating in these digital libraries has become very challenging.
- Keyphrases of a document can allow for efficient processing of more information in less time and improving many natural language processing and information retrieval tasks, e.g., summarization and contextual advertisement.
- Keyphrase extraction is defined as the problem of automatically extracting descriptive phrases or concepts from a document.

Previous Approaches to Keyphrase Extraction

- Many approaches to keyphrase extraction have been proposed in the literature along two lines of research: supervised and unsupervised.
- In the supervised line of research, different feature sets (e.g., term frequency, relative position of the first occurrence, part-of-speech tag) and classification algorithms (e.g., Naive Bayes) give rise to various supervised keyphrase extraction models.
- Intuitively, keyphrases occur very early in a document and appear frequently.

Proposed Approach

We propose a fully unsupervised graph-based algorithm that incorporates information from all positions of a word’s occurrences into a biased-PageRank to score keyphrases.

- Our approach involves three essential steps:
  1. The graph construction at the word level;
  2. The design of biased-PageRank algorithm;
  3. The scoring of multi-word phrases.

Graph Construction

Unsupervised Semantic Parsing

We present the first unsupervised approach to the problem of learning a semantic parser using Markov logic. Our USP system translates dependency trees into quasi-logical forms, recursively infers lambda forms from these, and clusters them to abstract away semantic variations of the same meaning. The MAP semantic parse of a sentence is obtained by recursively assigning to parts of lambda-form clusters and computing them. We evaluate our approach by using it to extract knowledge base from biomedical abstracts and answer questions. USP substantially surpasses TextRunner DBRT and an informed baseline on both precision and recall on this task.

- window = 3

Position Biased PageRank

- The idea of our approach is to assign higher probabilities to those words that occur very early in the document.
- We weight each candidate word with the inverse position in the document. If the same word appears multiple times in target document, then we add all its position weights.
- Similar to Haveliwala [Haveliwala(2002)], we biased PageRank to prefer these words by incorporating the weight of a word in the equation of PageRank as follows:

\[ p(v_i) = (1 - \alpha) \cdot p(v_i) + \alpha \cdot \sum_{j \in \text{neighbours}(v_i)} w_{ji} \cdot p(v_j) \]

- Multi-word phrases are scored by using the sum of scores of individual words that comprise the phrase [Wan & Xiao(2008)].

Datasets

- Datasets available at http://www.cse.unl.edu/~ccaragea/keyphrases.html

Baseline

- TF-IDF. Keyphrases are ranked based on their term-frequency-inverse document frequency score [Barker & Cornacchia(2001)].
- ExpandRank. A word graph was built for each paper and its local textual neighbors [Wan & Xiao(2008)].
- TopicalPageRank (TPR). Latent Dirichlet Allocation is used to infer the topic distribution of words and keyphrases are ranked by aggregating the topic-specific scores [Liu et al.(2010); Liu, Huang, Zheng, & Sun].

Results

How does our approach compare with other existing state-of-the-art methods?

- Conclusions: We proposed an unsupervised graph-based model which incorporates both the relative position and the frequency of a term into a biased PageRank.
- Our experiments on three datasets show that our proposed model achieves better performance than strong baselines.
- Future directions: Further evaluation of our approach on other types of documents, e.g., news articles, transcripts, etc.

References

[Barker & Cornacchia(2001)]
Using unsupervised biased random walk keyphrase extraction.
In Conference of the Canadian Society for Computational Studies of Intelligence, pp. 40-51.

[Caragea, C., Bulgarov, F., Godea, A., & Gollapalli, S. D. (2014)]
Cooccurrence-based keyphrase extraction as research papers. A supervised approach.

[Caragea, C., & Caragea, C. (2014)]
Extracting keyphrases from research papers using citation networks.

[Caragea, C., Bulgarov, F., Godea, A., & Gollapalli, S. D. (2014)]
Automatic keyphrase extraction via topic decomposition.

[Hulth, A. (2003)]
Topic-sensitive pagerank.
In Proceedings of the Twenty-Eighth Conference on Artificial Intelligence, pp. 1629-1635.

[Nguyen, H. (2015)]
Naive Bayes for keyphrase extraction.

[Textrank: Bringing order into text. (2010)]

Keyphrase Extraction: An Unsupervised Approach

Corina Florescu and Cornelia Caragea

Computer Science and Engineering, University of North Texas

Acknowledgements. This research is partially supported by a grant from the National Science Foundation to Corina Caragea.

WWW: https://corinaflorescu.github.io/cs