A Semi-Supervised Key Phrase Extraction Approach: Learning from Title Phrases through a Document Semantic Network

Decong Li[1], Sujian Li[1], Wenjie Li[2], Wei Wang[1], Weiguang Qu[3]
[1]Key Laboratory of Computational Linguistics, Peking University
[2]Department of Computing, The Hong Kong Polytechnic University
[3] School of Computer Science and Technology, Nanjing Normal University
{lidecong,lisujian, wwei }@pku.edu.cn cswjli@comp.polyu.edu.hk wgqu@njnu.edu.cn

Abstract

It is a fundamental and important task to extract key phrases from documents. Generally, phrases in a document are not independent in delivering the content of the document. In order to capture and make better use of their relationships in key phrase extraction, we suggest exploring the Wikipedia knowledge to model a document as a semantic network, where both *n*-ary and binary relationships among phrases are formulated. Based on a commonly accepted assumption that the title of a document is always elaborated to reflect the content of a document and consequently key phrases tend to have close semantics to the title, we propose a novel semi-supervised key phrase extraction approach in this paper by computing the phrase importance in the semantic network, through which the influence of title phrases is propagated to the other phrases iteratively. Experimental results demonstrate the remarkable performance of this approach.

1 Introduction

Key phrases are defined as the phrases that express the main content of a document. Guided by the given key phrases, people can easily understand what a document describes, saving a great amount of time reading the whole text. Consequently, automatic key phrase extraction is in high demand. Meanwhile, it is also fundamental to many other natural language processing applications, such as information retrieval, text clustering and so on.

Key phrase extraction can be normally cast as a ranking problem solved by either supervised or unsupervised methods. Supervised learning requires a large amount of expensive training data, whereas unsupervised learning totally ignores human knowledge. To overcome the deficiencies of these two kinds of methods, we propose a novel semi-supervised key phrase extraction approach in this paper, which explores title phrases as the source of knowledge.

It is well agreed that the title has a similar role to the key phrases. They are both elaborated to reflect the content of a document. Therefore, phrases in the titles are often appropriate to be key phrases. That is why position has been a quite effective feature in the feature-based key phrase extraction methods (Witten, 1999), i.e., if a phrase is located in the title, it is ranked higher.

However, one can only include a couple of most important phrases in the title prudently due to the limitation of the title length, even though many other key phrases are all pivotal to the understanding of the document. For example, when we read the title "China Tightens Grip on the Web", we can only have a glimpse of what the document says. On the other hand, the key phrases, such as "China", "Censorship", "Web", "Domain name", "Internet", and "CNNIC", etc. can tell more details about the main topics of the document. In this regard, title phrases are often good key phrases but they are far from enough.

If we review the above example again, we will find that the key phrase "Internet" can be inferred from the title phrase "Web". As a matter of fact, key phrases often have close semantics to title phrases. Then a question comes to our minds: can we make use of these title phrases to infer the other key phrases?

To provide a foundation of inference, a semantic network that captures the relationships among phrases is required. In the previous works (Turdakov and Velikhov, 2008), semantic networks are constructed based on the binary relations, and the semantic relatedness between a pair of phrases is formulated by the weighted edges that connects them. The deficiency of these approaches is the incapability to capture the *n*-ary relations among multiple phrases. For example, a group of

Proceedings of the ACL 2010 Conference Short Papers, pages 296–300,
Uppsala, Sweden, 11-16 July 2010.

phrases may collectively describe an entity or an event.

In this study, we propose to model a semantic network as a hyper-graph, where vertices represent phrases and weighted hyper-edges measure the semantic relatedness of both binary relations and *n*-ary relations among phrases. We explore a universal knowledge base – Wikipedia – to compute the semantic relatedness. Yet our major contribution is to develop a novel semi-supervised key phrase extraction approach by computing the phrase importance in the semantic network, through which the influence of title phrases is propagated to the other phrases iteratively.

The goal of the semi-supervised learning is to design a function that is sufficiently smooth with respect to the intrinsic structure revealed by title phrases and other phrases. Based on the assumption that semantically related phrases are likely to have similar scores, the function to be estimated is required to assign title phrases a higher score and meanwhile locally smooth on the constructed hyper-graph. Zhou et al.'s work (Zhou 2005) lays down a foundation for our semi-supervised phrase ranking algorithm introduced in Section 3. Experimental results presented in Section 4 demonstrate the effectiveness of this approach.

2 Wikipedia-based Semantic Network Construction

Wikipedia[1] is a free online encyclopedia, which has unarguably become the world's largest collection of encyclopedic knowledge. *Articles* are the basic entries in the Wikipedia, with each article explaining one Wikipedia term. Articles contain *link*s pointing from one article to another. Currently, there are over 3 million articles and 90 million links in English Wikipedia. In addition to providing a large vocabulary, Wikipedia articles also contain a rich body of lexical semantic information expressed via the extensive number of links. During recent years, Wikipedia has been used as a powerful tool to compute semantic relatedness between terms in a good few of works (Turdakov 2008).

We consider a document composed of the phrases that describe various aspects of entities or events with different semantic relationships. We then model a document as a semantic network formulated by a weighted hyper-graph $G=(V, E, W)$, where each vertex $v_i \in V$ $(1 \le i \le n)$ represents a phrase, each hyper-edge $e_j \in E$ $(1 \le j \le m)$ is a subset of V, representing binary relations or *n*-ary relations among phrases, and the weight $w(e_j)$ measures the semantic relatedness of e_j.

By applying the WSD technique proposed by (Turdakov and Velikhov, 2008), each phrase is assigned with a single Wikipedia article that describes its meaning. Intuitively, if the fraction of the links that the two articles have in common to the total number of the links in both articles is high, the two phrases corresponding to the two articles are more semantically related. Also, an article contains different types of links, which are relevant to the computation of semantic relatedness to different extent. Hence we adopt the weighted Dice metric proposed by (Turdakov 2008) to compute the semantic relatedness of each binary relation, resulting in the edge weight $w(e_{ij})$, where e_{ij} is an edge connecting the phrases v_i and v_j.

To define the *n*-ary relations in the semantic network, a proper graph clustering technique is needed. We adopt the weighted Girvan-Newman algorithm (Newman 2004) to cluster phrases (including title phrases) by computing their betweenness centrality. The advantage of this algorithm is that it need not specify a pre-defined number of clusters. Then the phrases, within each cluster, are connected by a *n*-ary relation. *n*-ary relations among the phrases in the same cluster are then measured based on binary relations. The weight of a hyper-edge e is defined as:

$$w(e) = \frac{\alpha}{|e|} \sum_{e_{ij} \subseteq e} w(e_{ij}) \quad (1)$$

where $|e|$ is the number of the vertices in e, e_{ij} is an edge with two vertices included in e and $\alpha \ge 0$ is a parameter balancing the relative importance of *n*-ary hyper-edges compared with binary ones.

3 Semi-supervised Learning from Title

Given the document semantic network represented as a phrase hyper-graph, one way to make better use of the semantic information is to rank phrases with a semi-supervised learning strategy, where the title phrases are regarded as labeled samples, while the other phrases as unlabeled ones. That is, the information we have at the beginning about how to rank phrases is that the title phrases are the most important phrases. Initially, the title phrases are assigned with a positive score of 1 indicating its importance and oth-

[1] www.wikipedia.org

er phrases are assigned zero. Then the importance scores of the phrases are learned iteratively from the title phrases through the hyper-graph. The key idea behind hyper-graph based semi-supervised ranking is that the vertices which usually belong to the same hyper-edges should be assigned with similar scores. Then, we have the following two constraints:

1. The phrases which have many incident hyper-edges in common should be assigned similar scores.

2. The given initial scores of the title phrases should be changed as little as possible.

Given a weighted hyper-graph G, assume a ranking function f over V, which assigns each vertex v an importance score $f(v)$. f can be thought as a vector in Euclid space $R^{|V|}$. For the convenience of computation, we use an incidence matrix H to represent the hypergraph, defined as:

$$h(v,e)=\begin{cases}0, & \text{if } v\notin e\\ 1, & \text{if } v\in e\end{cases} \quad (2)$$

Based on the incidence matrix, we define the degrees of the vertex v and the hyper-edge e as

$$d(v)=\sum_{e\in E} w(e)h(v,e) \quad (3)$$

and

$$\delta(e)=\sum_{v\in V} h(v,e) \quad (4)$$

Then, to formulate the above-mentioned constraints, let y denote the initial score vector, then the importance scores of the phrases are learned iteratively by solving the following optimization problem:

$$\arg\min_{f\in R^{|V|}}\{\Omega(f)+\mu\|f-y\|^2\} \quad (5)$$

$$\Omega(f)=\frac{1}{2}\sum_{e\in E}\frac{1}{\delta(e)}\sum_{\{u,v\}\subseteq e} w(e)\left(\frac{f(u)}{\sqrt{d(u)}}-\frac{f(v)}{\sqrt{d(v)}}\right)^2 \quad (6)$$

where $\mu>0$ is the parameter specifying the tradeoff between the two competitive items. Let D_v and D_e denote the diagonal matrices containing the vertex and the hyper-edge degrees respectively, W denote the diagonal matrix containing the hyper-edge weights, f^* denote the solution of (6). Zhou has given the solution (Zhou, 2005) as.

$$f^*=\beta\Theta f^*+(1-\beta)y \quad (7)$$

where $\Theta=D_v^{-1/2}HWD_e^{-1}H^TD_v^{-1/2}$ and $\beta=1/(\mu+1)$. Using an approximation algorithm (e.g. Algorithm 1), we can finally get a vector f representing the approximate phrase scores.

Algorithm 1: PhraseRank(V, T, a, b)

Input: Title phrase set = $\{v_1,v_2,\ldots,v_t\}$,the set of other phrases =$\{v_{t+1},v_{t+2},\ldots,v_n\}$, parameters α and β, convergence threshold ζ

Output: The approximate phrase scores f

Construct a document semantic network for all the phrases $\{v_1,v_2,\ldots,v_n\}$ using the method described in section 2.

Let $\Theta=D_v^{-1/2}HWD_e^{-1}H^TD_v^{-1/2}$;

Initialize the score vector y as $y_i=1, 1\le i\le t$, and $y_j=0, t<j\le n$;

Let $f^0=y$, $k=0$;

REPEAT

$f^{k+1}=\beta\Theta f^k+(1-\beta)y$;

$\nabla\leftarrow max|f_i^{k+1}-f_i^k|$, for $0\le i\le n$;

$k\leftarrow k+1$;

UNTIL $\nabla<\zeta$

END

Finally we rank phrases in descending order of the calculated importance scores and select those highest ranked phrases as key phrases. According to the number of all the candidate phrases, we choose an appropriate proportion, i.e. 10%, of all the phrases as key phrases.

4 Evaluation

4.1 Experiment Set-up

We first collect all the Wikipedia terms to compose of a dictionary. The word sequences that occur in the dictionary are identified as phrases. Here we use a finite-state automaton to accomplish this task to avoid the imprecision of preprocessing by POS tagging or chunking. Then, we adopt the WSD technique proposed by (Turdakov and Velikhov 2008) to find the corresponding Wikipedia article for each phrase. As mentioned in Section 2, a document semantic network in the form of a hyper-graph is constructed, on which Algorithm 1 is applied to rank the phrases.

To evaluate our proposed approach, we select 200 pieces of news from well-known English media. 5 to 10 key phrases are manually labeled in each news document and the average number of the key phrases is 7.2 per document. Due to the abbreviation and synonymy phenomena, we construct a thesaurus and convert all manual and automatic phrases into their canonical forms when evaluated. The traditional Recall, Precision and F1-measure metrics are adopted for evaluation. This section conducts two sets of experiment: (1) to examine the influence of two parameters: α and β, on the key phrase extraction performance; (2) to compare with other well known state-of-art key phrase extraction approaches.

4.2 Parameter tuning

The approach involves two parameters: α ($\alpha \geq 0$) is a relation factor balancing the influence of *n*-ary relations and binary relations; β ($0 \leq \beta \leq 1$) is a learning factor tuning the influence from the title phrases. It is hard to find a global optimized solution for the combination of these two factors. So we apply a gradient search strategy. At first, the learning factor is set to β=0.8. Different values of α ranging from 0 to 3 are examined. Then, given that α is set to the value with the best performance, we conduct experiments to find an appropriate value for β.

4.2.1 α: Relation Factor

First, we fix the learning factor β as 0.8 randomly and evaluate the performance by varying α value from 0 to 3. When α=0, it means that the weight of n-ary relations is zero and only binary relations are considered. As we can see from Figure 1, the performance is improved in most cases in terms of F1-measure and reaches a peak at α=1.8. This justifies the rational to incorporate *n*-ary relations with binary relations in the document semantic network.

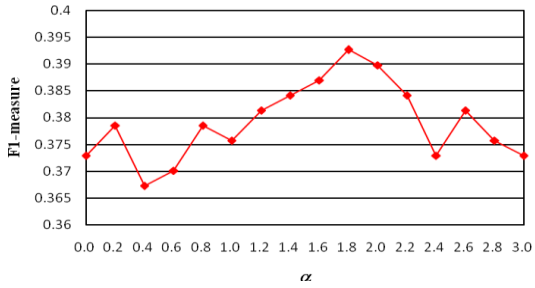

Figure 1. F1-measures with α in [0 3]

4.2.2 β: Learning factor

Next, we set the relation factor α=1.8, we inspect the performance with the learning factor β ranging from 0 to 1. β=1 means that the ranking scores learn from the semantic network without any consideration of title phrases. As shown in Figure 2, we find that the performance almost keep a smooth fluctuation as β increases from 0 to 0.9, and then a diving when β=1. This proves that title phrases indeed provide valuable information for learning.

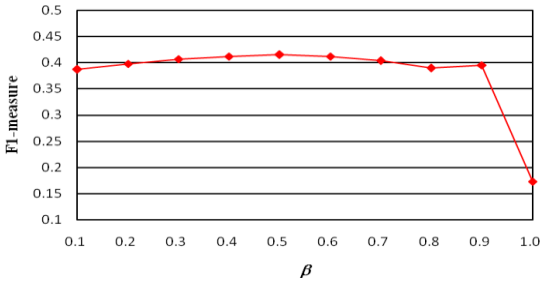

Figure 2. F1-measure with β in [0,1]

4.3 Comparison with Other Approaches

Our approach aims at inferring important key phrases from title phrases through a semantic network. Here we take a method of synonym expansion as the baseline, called WordNet expansion here. The WordNet[2] expansion approach selects all the synonyms of the title phrases in the document as key phrases. Afterwards, our approach is evaluated against two existing approaches, which rely on the conventional semantic network and are able to capture binary relations only. One approach combines the title information into the Grineva's community-based method (Grineva *et al.*, 2009), called title-community approach. The title-community approach uses the Girvan-Newman algorithm to cluster phrases into communities and selects those phrases in the communities containing the title phrases as key phrases. We do not limit the number of key phrases selected. The other one is based on topic-sensitive LexRank (Otterbacher *et al.*, 2005), called title-sensitive PageRank here. The title-sensitive PageRank approach makes use of title phrases to re-weight the transitions between vertices and picks up 10% top-ranked phrases as key phrases.

Approach	Precision	Recall	F1
Title-sensitive PageRank (d=0.15)	34.8%	39.5%	37.0%
Title-community	29.8%	**56.9%**	39.1%
Our approach (α=1.8, β=0.5)	**39.4%**	44.6%	**41.8%**
WordNet expansion (baseline)	7.9%	32.9%	12.5%

Table 1. Comparison with other approaches

Table 1 summarizes the performance on the test data. The results presented in the table show that our approach exhibits the best performance among all the four approaches. It follows that the key phrases inferred from a document semantic network are not limited to the synonyms of title phrases. As the title-sensitive PageRank ap-

[2] http://wordnet.princeton.edu

proach totally ignores the *n*-ary relations, its performance is the worst. Based on binary relations, the title-community approach clusters phrases into communities and each community can be considered as an *n*-ary relation. However, this approach lacks of an importance propagation process. Consequently, it has the highest recall value but the lowest precision. In contrast, our approach achieves the highest precision, due to its ability to infer many correct key phrases using importance propagation among *n*-ary relations.

5 Conclusion

This work is based on the belief that key phrases tend to have close semantics to the title phrases. In order to make better use of phrase relations in key phrase extraction, we explore the Wikipedia knowledge to model one document as a semantic network in the form of hyper-graph, through which the other phrases learned their importance scores from the title phrases iteratively. Experimental results demonstrate the effectiveness and robustness of our approach.

Acknowledgments
The work described in this paper was partially supported by NSFC programs (No: 60773173, 60875042 and 90920011), and Hong Kong RGC Projects (No: PolyU5217/07E). We thank the anonymous reviewers for their insightful comments.

References

David Milne, Ian H. Witten. 2008. *An Effective, Low-Cost Measure of Semantic Relatedness Obtained from Wikipedia Links*. In Wikipedia and AI workshop at the AAAI-08 Conference, Chicago, US.

Dengyong Zhou, Jiayuan Huang and Bernhard Schölkopf. 2005. *Beyond Pairwise Classification and Clustering Using Hypergraphs*. MPI Technical Report, Tübingen, Germany.

Denis Turdakov and Pavel Velikhov. 2008. *Semantic relatedness metric for wikipedia concepts based on link analysis and its application to word sense disambiguation*. In Colloquium on Databases and Information Systems (SYRCoDIS).

Ian H. Witten, Gordon W. Paynter, Eibe Frank , Carl Gutwin , Craig G. Nevill-Manning. 1999. *KEA: practical automatic keyphrase extraction*, In Proceedings of the fourth ACM conference on Digital libraries, pp.254-255, California, USA.

Jahna Otterbacher, Gunes Erkan and Dragomir R. Radev. 2005. *Using Random Walks for Question-focused Sentence Retrieval*. In Proceedings of HLT/EMNLP 2005, pp. 915-922, Vancouver, Canada.

Maria Grineva, Maxim Grinev and Dmitry Lizorkin. 2009. *Extracting key terms from noisy and multitheme documents*, In Proceedings of the 18th international conference on World wide web, pp. 661-670, Madrid, Spain.

Michael Strube and Simone Paolo Ponzetto. 2006.WikiRelate! *Computing Semantic Relatedness using Wikipedia.* In Proceedings of the 21st National Conference on Artificial Intelligence, pp. 1419–1424, Boston, MA.

M. E. J. Newman. 2004. *Analysis of Weighted Networks*. Physical Review E 70, 056131.