Hypergraph-based Inductive Learning for Generating Implicit Key Phrases

Decong Li Sujian Li
Key Laboratory of Computational Linguistics, Ministry of Education, CHINA
Institute of Computational Linguistics, Peking University, Beijing, 100871, China
Tel: 86-10-62753081-105
{lidecong, lisujian}@pku.edu.cn

ABSTRACT
This paper presents a novel approach to generate implicit key phrases which are ignored in previous researches. Recent researches prefer to extract key phrases with semi-supervised transductive learning methods, which avoid the problem of training data. In this paper, based on a transductive learning method, we formulate the phrases in the document as a hypergraph and expand the hypergraph to include implicit phrases, which are ranked by an inductive learning approach. The highest ranked phrases are seen as implicit key phrases, and experimental results demonstrate the satisfactory performance of this approach.

Categories and Subject Descriptors
H.3.1 **[Information Storage and Retrieval]**: Content Analysis and Indexing – *indexing methods.*

General Terms
Algorithms, Experimentation.

Keywords
Key phrase generation, Implicit key phrase, Transductive learning, Inductive semi-supervised learning, Hypergraph.

1. INTRODUCTION
Key phrases (or key terms, keywords) are the phrases which deliver the main content of a document. Most state-of-the-art researches generate key phrases through extracting directly from the document itself. However, some phrases contribute greatly to understanding the document though they do not appear in the text. These phrases are also appropriate to be key phrases, as a beneficial complement of the extracted ones, called as *implicit key phrases* in this paper. Implicit key phrases are important to many research fields such as information retrieval, information extraction and advertising.

To our knowledge, so far the task of implicit key phrases generation is often implemented as a multiclass classification problem with the help of some external resources, such as controlled vocabularies [1] or conceptual thesaurus [2]. A drawback of the classification methods is that they require a large manually annotated corpus, as a suitable number of training examples are needed for each possible implicit key phrase. To solve the problem of training data, transductive learning methods have been proposed. It is well agreed that all the phrases in one document have semantic relatedness which can be formulated with a graph. Still, the title of a document usually includes a couple of important phrases. With the commonsense above, Li [3] proposes a semi-supervised transductive learning algorithm for key phrase extraction, which learns the phrase importance from the title phrases through a hypergraph. Furthermore, one problem comes to our minds: can one document be self-explained by the phrases in the document? As we know, one document is often compiled with some pertinent knowledge omitted, which may contribute a lot to the understanding and retrieval. Here, we expect to use Wikipedia, the free online encyclopedia, to expand the content of the document to include more pertinent knowledge, represented by the related implicit phrases.

Next it is the problem how the implicit phrases are ranked. Since the extracted phrases express the main content and the implicit phrases can be seen as a helpful complement, the importance of implicit phrases should be ranked mainly based on the extracted phrases. Here, we make full use of the transductive semi-supervised learning results which rank the extracted phrases through a hypergraph. The assumption is: the higher ranked the extracted phrases which an implicit phrase is related to are, the higher ranked the implicit phrase. It is fortunately that Delalleau's [4] algorithm caters to our assumption, under which a graph-based inductive learning method can rank previously unseen test data through generalizing the transductive learning on the training data (labeled and unlabeled). Based on Delalleau's model, we propose a novel hypergraph-based inductive learning approach to rank implicit phrases, whose difference with Delalleau's is that a hypergraph can contain more abundant relations of phrases, including binary and *n*-ary relations.

2. HYPERGRAPH-BASED INDUCTIVE LEARNING APPOACH
Our approach is based on the transductive learning method [3], where all extracted phrases compose a hypergraph and learn their importance iteratively through the hypergraph with title phrases labeled important. Then, a document is formulated by a weighted hypergraph $G=(V, E, W)$, where each vertex $v_i \in V$ $(1 \le i \le n)$ represents a phrase, each hyperedge $e_j \in E$ $(1 \le j \le m)$ 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. The key idea behind hypergraph based transductive ranking is that the vertices which usually belong to the same hyperedges should be assigned with similar scores. Given such a weighted hypergraph G, assume a ranking function f over V, which assigns each vertex v an importance score $f(v)$. f can be cast as a vector in Euclid space $R^{|V|}$ with the minimization criterion:

$$\arg\min_{f \in R^{|V|}} \{\Omega(f) + \mu \|f - y\|^2\} \quad (1)$$

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

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0637-9/11/03.

where $\delta(e)$ and $d(v)$ denote the degrees of hyperedge e, and vertex v respectively, y denote the initial score vector.

Since the implicit phrases are just a complement of the document, they should not change much of the original hypergraph, which have reached a stable state during the transductive learning. The reasonable assumption is that the value of the function over the existing phrases will not change with the addition of a new phrase. Adding terms for a new implicit phrase x in Eq(1) and keeping the value of f fixed on the extracted phrases lead to the minimization of the modified criterion:

$$\operatorname{argmin}_{f(x)} \sum_{e \in E'} \frac{1}{\delta(e)} \sum_{\{x,v\} \subseteq e} w(e) \left(\frac{f(x)}{\sqrt{d(x)}} - \frac{f(v)}{\sqrt{d(v)}} \right)^2$$

This function above is convex in $f(x)$ and is minimized when

$$f^*(x) = \frac{\sum_{e \in E'} \frac{1}{\delta(e)} \sum_{\{x,v\} \subseteq e} w(e) \frac{f(v)\sqrt{d(x)}}{\sqrt{d(v)}}}{\sum_{e \in E'} \frac{1}{\delta(e)} \sum_{\{x,v\} \subseteq e} w(e)} \quad (2)$$

The foundation of this work is the Wikipedia knowledge. After a hypergraph is constructed and the transductive learning is conducted for the extracted phrases, implicit phrases are generated according to the document. Since the newly added implicit phrases should be highly related to the document and the title phrases are always elaborated to reflect the content of a document, we choose the hyperlink phrases from the corresponding Wikipedia articles explaining the title phrases and include them in the hypergraph. In the module of hypergraph expansion, we need construct the semantic relations between the newly added phrase and the other phrases, including binary and n-ary relations represented by the weighted hyperedges set E'. Binay relations are computed by adopting weighted *Dice* metric as in [5]. To acquire the new n-ary relations, we employ a classification strategy, which first calculates the *Dice* metric between the new phrase and each hyperedge (a group of phrases) and then classify the phrase into the hyperedge with the highest *Dice* score. Of course, the weight of the changed hyperedge need to be modified, re-computed as in [3]. Finally, according to Eq (2) we compute $f^*(x)$ for each candidate phrases and select the top ranked ones as the implicit key phrases.

3. EXPERIMENTS

We use two datasets to evaluate our approach: one dataset is composed of 100 pieces of news, which we collect from well-known English media; the other one contains 50 scientific articles from Task 5 (named: Automatic keyphrase extraction from scientific articles) of SemEva-2[1].

To our knowledge, there is no uniform metric to evaluate implicit key phrases. To evaluate our approach, we propose to set three kinds of ranking levels for each implicit phrase: (a) fit for key phrase (*KP*), (b) not fit enough but closely related (*RL*) and (c) not fit at all (*NF*). For each test document, we ask several well-educated human evaluators majoring in economics, politics, and computer science to assign a ranking level to each implicit key phrase generated using different methods.

Since an intuitive approach to generate implicit phrases is to use a thesaurus, here we take a method of WordNet[2] based expansion as the baseline, which selects all the *synonym*s and the nearest-level *hypernym*s (if they have) of the title phrases in the document as key phrases. The ones that do not appear in the document are seen as implicit key phrases. Our approach is also evaluated against the semi-supervised transductive learning methods (HTL) adopted by Li [3], which now applies on a hypergraph composed of both the extracted phrases and the implicit phrases. For our inductive learning approach (OURS) and the transductive learning approach (HTL), the number of the implicit phrases is controlled around 10 for each document. The WordNet based approach (WORDNET) cannot control the size of the implicit key phrase set and generates about 28.3 implicit key phrases on average for each document.

Table 1. Comparison of two approaches on two test sets

Approach	News (%)			Scientific Articles (%)		
	KP	*RL*	*NF*	*KP*	*RL*	*NF*
OURS	**35.7**	48.7	15.6	**31.0**	43.0	26.0
HTL	17.5	37.5	45.0	22.5	40.0	37.5
WORDNET	22.7	**53.5**	23.8	21.1	**47.3**	31.6

Table 1 summarizes the performance on the two test datasets. We can see that our approach has the highest proportion of implicit key phrases ranked as *KP*, the WordNet-based approach has the highest proportion of implicit key phrases ranked as *RL*, and the HTL approach performs the worst of the three. The disadvantage of WordNet based approach is that it collects all the synonyms and hypernyms, but cannot further differentiate these phrases which are appropriate to be key phrases. Then most generated implicit phrases are somewhat related to the document, and only a small portion can be seen as key phrases. The reason that HTL does not surpass the baseline is also easy to explain. In this approach, the implicit phrases have the same status with the extracted phrases, and this may introduce more noise into the computing process because most implicit phrases are far from the content of the document. Our approach overcomes the drawbacks of these two approaches. On one hand, it can make full use of the Wikipedia to judge the meaning of each implicit phrase. On the other hand, our approach ranks each implicit phrase in parallel, meaning implicit phrases do not influence each other.

4. REFERENCES

[1] O. Medelyan and I. H. Witten. *Thesaurus based automatic keyphrase indexing*. In: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries, pp. 296–297. ACM Press, New York, 2006.

[2] B. Pouliquen, R. Steinberger and C. Ignat. *Automatic annotation of multilingual text collections with a conceptual thesaurus*. In: BUG 2003.

[3] D. Li, S. Li and W. Li. *A Semi-Supervised Key Phrase Extraction Approach: Learning from Title Phrases through a Document Semantic Network*. In: Proceedings of the 48thACL, pp 296-300, Uppsala, Sweden, 2010.

[4] O. Delalleau, Y. Bengio and N. Le Roux., *Efficient Non-Parametric Function Induction in Semi-Supervised Learning*. Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics, pp 96-103, UK, Jan 6-8, 2005.

[5] D. Turdakov and P. Velikhov. *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), 2008.

[1] http://semeval2.fbk.eu/semeval2.php?location=tasks

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