Expert Systems with Applications 36 (2009) 9333–9339

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Using ontology to improve precision of terminology extraction from documents

Wen Zhang [a,c,*], Taketoshi Yoshida [a], Xijin Tang [b]

[a] School of Knowledge Science, Japan Advanced Institute of Science and Technology, 1-1 Asahidai Tatsunokuchi, Ishikawa 923-1292, Japan
[b] Institute of Systems Science, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, PR China
[c] Information Systems Department, School of Economics and Management, Beihang University, Beijing 100083, PR China

ARTICLE INFO

Keywords:
Terminology extraction
Ontology
Semantic dependency
WordNet

ABSTRACT

In this paper, we proposed a new approach using ontology to improve precision of terminology extraction from documents. Firstly, a linguistic method was used to extract the terminological patterns from documents. Then, similarity measures within the framework of ontology were employed to rank the semantic dependency of the noun words in a pattern. Finally, the patterns at a predefined proportion according to their semantic dependencies were retained and regarded as terminologies. Experiments on Retuers-21578 corpus has shown that WordNet ontology, that we adopted for the task of extracting terminologies from English documents, can improve the precision of classical linguistic method on terminology extraction significantly.

1. Introduction

Recently, terminology and its related lexical units such as multi-words, collocations, etc. have been widely studied in both linguistics and text mining field. From the cognitive point of view, human beings recognize, learn and understand the entities and concepts in texts for a complete natural language comprehension. It is commonly accepted by researchers in natural language processing (NLP) that terminologies can better capture the topics of texts and describe the contents of texts more accurate than individual words, because their distinctive entities in a domain and their referents are more specific and unambiguous than their constituents as individual words where polysemy may usually occur.

Although "terminology" is the fundamental notion of this paper, this notion has no satisfactory formal definition currently. Actually, terminology has some overlappings with noun phrase and collocation, so it is not easy to differentiate them in definition precisely. For instance, they all contain two or more individual words in word-building and these words are adjacent to each other in a sequence. However, they have different intentions to characterize the lexical units in texts. Based on Samadja's definition (Smadja, 1993), there are three types of collocations: predictive relations, rigid noun phrases and phrasal templates. For instance, "his book" is a non phrase but only when it occurs frequently (statistical characteristics) in texts can it be regarded as a collocation. Terminology has more expert meaning in a specific domain. For instance, "normal distribution" is a terminology in mathematical domain, but mostly, we do not regard "his book" as a terminology.

There are mainly two types of methods developed for terminology extraction: linguistic method and statistical method. Linguistic methods utilize structural properties of phrases and sentence grammar of a special language to extract terminologies from documents (Chen & Chen, 1994; Choueka, 1983; Church, 1989; Justeson & Katz, 1995). Statistical methods utilize corpus learning with statistical indicators to measure the words' association for co-occurrence pattern discovery.

In the linguistic aspect, Choueka's methodology for handling the large corpora can be considered as a first step toward computer-aided lexicography. In his method, the consecutive sequences with two to six words were retrieved as collocations (Choueka, 1983). Justeson and Katz proposed a regular expression for individual words in a consecutive sequence to retrieve terminology (Justeson & Katz, 1995). We will discuss his work in details later in Section 2.2. Chen and Chen utilized the linguistic knowledge to extract the noun phrases by a finite state mechanism. They reported that their method can produce a recall as 95% and precision as 96% at average on all the categories of SUSANNE corpus (online: http://www.grsampson.net/Resources.html) (Chen & Chen, 1994).

In the statistical aspect, language is modeled as a stochastic process and the corpus is used to estimate that whether or not a given sequence occurs in the corpus by chance. Church and Hanks proposed the association ratio for measuring word association based on the information theoretic concept of mutual information to retrieve the pairs of words which occurred frequently in corpus together (Church & Hanks, 1990). Smadja developed Xtract to extract collocations from documents using the relative positions of two words in a corpus (Smadja, 1993). In Xtract, four parameters were used: strength, spread, peak z-score and percentage frequency.

* Corresponding author. Address: School of Knowledge Science, Japan Advanced Institute of Science and Technology, 1-1 Asahidai Tatsunokuchi, Ishikawa 923-1292, Japan.
E-mail addresses: zhangwen@jaist.ac.jp (W. Zhang), yoshida@jaist.ac.jp (T. Yoshida), xjtang@amss.ac.cn (X. Tang).

0957-4174/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2008.12.034

Taking a word pair (X, Y) for example, strength is used to describe the relative occurrence of X and Y to the occurrence of X and other words than Y. Spread is used to describe the variance of the relative positions of Y to X. Peak z-score is a given threshold which is used to filter out the words co-occurred with X but their relative positions to X are not kept stable. Percentage frequency is also a threshold used to define the percentage above which the word pairs with highest frequencies were regarded as collocations (Smadja, 1993).

However, concerning the methods mentioned above, they only regarded words in a sequence as individual characters other than meaningful lexical unit, especially in the statistical methods. Although part of speeches of individual words in a sequence was considered in linguistic methods, it is not enough to capture specific concepts from texts because more importantly, individual words constitute a terminology is in that they have some semantic relationship in essence. For this reason, we argue that for terminology extraction, especially in the case that we want to extract some domain related concepts from documents, the semantic relationship between individual words in a sequence should be emphasized particularly.

The contribution of this paper is threefold. Firstly, ontology, WordNet and the relationships of words within ontology is introduced. Secondly, a method is proposed to rank words' semantic dependency in a sequence based on word similarities within ontology. Thirdly, a classical linguistic method is employed to extract word patterns from an English corpus and the proposed method is examined.

The reminder of this paper is organized as follows. In Section 2, Justeson and Katz's linguistic approach for terminology extraction is reviewed and some basic ideas of ontology and the semantic relationship of words in WordNet are introduced. In Section 3, similarity of words within the framework of ontology is discussed and a method to rank the words' semantic dependency is proposed. In Section 4, a series of experiments on terminology extraction from Reuters-21578 t are conducted o validate the effectiveness of the proposed method. Section 5 concludes this paper and indicates our future work.

2. Preliminaries

A classic linguistic method for terminology extraction was presented and the semantic relationship of noun words within ontology was discussed in this section.

2.1. A linguistic approach for terminology extraction

From Justeson and Katz's point of view, noun phrases (NP) can be divided into to two groups: lexical NPs and non-lexical NPs. Lexical NPs are subject to a much more restricted range and extent of modifier variation, on repeated references to the entities they designate, than non-lexical NPs. And the terminological NPs differ from other NPs because they are lexical (Justeson & Katz, 1995).

When a terminological NP is a topic of significant discussion within a text, they tend to be repeated intact on repeated references to the entities they designate. The non-lexical NPs usually do not repeat many times within a text because they can simply be represented by the head noun and their modifiers often vary. For this reason, one effective criterion for terminology identification is simple repetition: a noun phrase having a frequency of two or more can be entertained as a likely terminological unit, i.e., as a candidate for inclusion in a list of technical terms from documents.

On the other hand, it is widely accepted that both lexical NPs and non-lexical NPs have the common characteristics as with length as two to six words and ending as a noun. It is also recognized that terminological NPs differ in structure, at least statistically, from non-lexical NPs. Experiments in English corpora showed that 97% of terminologies consists of nouns and adjectives only, and more than 99% consist only of nouns, adjectives and the preposition as "of".

Based on the above analysis concerning the characteristics of terminologies in documents, Justeson and Katz proposed two constraints to identify the terminologies in documents (referred as it JK method hereafter).

1. Frequency: Candidates strings must have frequency of 2 or more in the text.
2. Grammatical structure: Candidate strings are those multi-word noun phrases that are specified by the regular expression $((A|N)^+|((A|N)^* (NP)^2)(A|N)^*)N$, where A is an adjective, N is a noun and P is a preposition.

For a pattern of length L, there are $(L+2)\cdot 2^{L-3}$ types of terminologies. They reported that their method can obtain recall as high as 97% and at least 77% precision in identifying the terminologies in documents, and at least 67% terminology types were conformed to the regulation given by them.

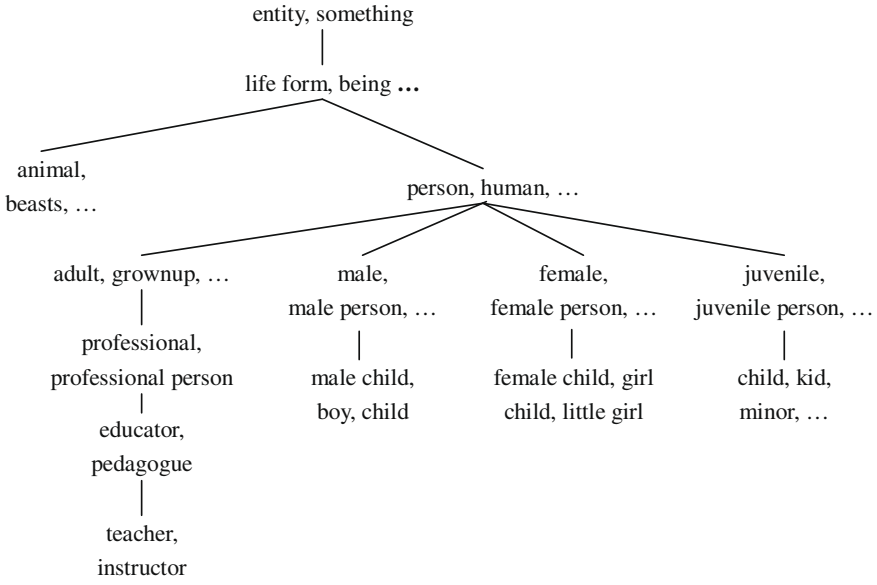

Fig. 1. Hierarchical semantic knowledge base. "..." indicates that some words in the class were omitted to save space (Li, 2003).

2.2. Ontology

In philosophy, ontology is a study of being or existence and forms the basic subject matter of metaphysics. It seeks to describe the basic categories and relationships of being or existence to define entities and types of entities within its framework (Ontology). In the area of knowledge management, ontology refers to using hierarchical trees to represent the background knowledge, for example, MESH ontology (online: http://www.nlm.nih.gov/mesh/) and WordNet (online: http://wordnet.princeton.edu/). Although no formal definition of ontology is generally recognized and how it should be implemented is controversial, we adopted the same definition as in Kohler (2006) for practical use in this paper. That is, ontology is constructed based on a controlled vocabulary and the relationships of the concepts in the controlled vocabulary.

Definition 1. Controlled vocabulary CV = name set of concepts c with c = (name, definition, identifier, synonyms).

In ontologies the concepts are linked by directed edges, then form a graph. The edges of an ontology specify in which way concepts are related to each other, e.g., "is-a" or "part-of".

Definition 2. Ontology $O = G$ (CV, E) with $E \subseteq CV \times CV$ and a totally defined function t: $E \rightarrow T$, which defines the types of edges. T is the of possible edge types, i.e., the semantics of an edge in natural language.

Fig. 1 is a segment of the WordNet ontology which is strictly constructed according to the above definition. From this figure, we can see that a concept in the hierarchical is represented by a set of synonymy and the "is a" relationships between concepts are represented by the edges connecting these concepts.

2.3. Semantic relationship of words in WordNet

WordNet try to make the semantic relations between word senses more explicit and easier to use. Because terminologies are usually nouns, in this paper, we concentrate on using noun words in WordNet to improve the performance of terminology extraction. WordNet (version 1.5) contains 80,000 noun word forms organized into some 60,000 lexicalized concepts. Many of these nouns are collocations; a few are artificial collocations invented for the convenience of categorization. WordNet divided the nouns into 11 hierarchies, each with a different unique beginner which corresponds to a primitive semantic component in a compositional theory of lexical semantics (Miller, 1998). The basic relationship in WordNet is synonymy. A set of synonym is called a synset. And the relationship between noun synsets in WordNet is either hypernym or hyponym. For instance, the synset "person, human" is a hypernym of the concept as "adult, grownup" and the relationship is hyponym in reverse. A synset has only one hypernym but it may have more than one hyponyms. This design for concepts in WordNet is very similar to the concept organization in human natural language. The distinctiveness of WordNet from conventional dictionary is that we can use the semantic relationships between synsets for inferences besides it is readable by computer. For instance, if we have a concept as "human", then we can infer that perhaps this "human" is "male" in gender and a "teacher" in vocation. More details concerning the nouns in WordNet can refer to Miller (1998).

In practical application, Rodriguez et al. used WordNet as additional lexical database to increase the amount of information in Vector Space Model for the task of text categorization (TC) on Reuters-21578 text collection. They reported that the integration of WordNet clearly improved the performance of Rocchio and Widrow-Hoff algorithms in TC (Rodriguez et al., 2000). Scott and Matwin developed the hypernym density representation using WordNet hypernyms and conducted TC using Ripper system. Their results showed that for some of the more difficult tasks their new representation method leads to significantly more accurate and more comprehensive rules (Scott & Matwin, 1998). Hotho et al. developed different strategies to compile the background knowledge embodied in ontologies into text documents representation. They reported that domain specific ontology can improve the clustering performance more significantly than general ontoloty (Hotho, 2003). Zhou et al proposed a new context-sensitive smoothing method in information retrieval (IR) which decomposes a document into a set of weighted context-sensitive topic signatures and then maps those topic signatures into query terms. Ontology was used as the topic signatures in their method. Experiments on TREC 2004/2005 Genomics Track (online: http://trec.nist.gov/data.html) demonstrate that ontologies used as topic signatures can significantly improve the IR performance over the traditional language model (Zhou, 2007). More work on using ontology for intelligent information processing can be found in Kohler (2006), Bloehdorn et al. (2004), Li (2008), Correa and Ludermir (2006).

3. Using ontology to improve terminology extraction

In this section, the motivation of adopting ontology into terminology extraction is specified. Two methods for similarity measure of words within ontology are described. The new method of combing ontology into terminology extraction is proposed.

3.1. The motivation

The main motivation to adopt ontology for terminology extraction is that, we want to make use of the background knowledge and words' semantic relationships compiled in ontology to capture the semantic features of terminologies in documents. We conjecture that ontology will take a positive effect on terminology extraction based on the following reasons:

1. Terminology is expression of a specific concept in a domain and ontology is also constructed on different domains. For this reason, we can use the concepts in ontology to match the terminologies in documents directly. For example, we can extract "professional person" directly from Fig. 1 and regard it as a terminology in "human" domain.
2. The constituents as individual words of a terminology are highly semantically correlated with each other and most lexical noun phrases have the property as compositional meaning in its sense. For this reason, we can deduce that the individual words in a terminology are of closer semantic relationships than those individual words not in a terminology but co-occurred together. For instance, "professional educator" is a terminology and the individual words "professional" and "educator" has close semantic relationship in WordNet ontology as shown in Fig. 1.
3. In document writing, terminologies are fixed phrases and its constituents co-occur together to express a complete concept. In WordNet ontology, all senses of a word are listed to relate its subordinates and superordinates. Although some words co-occur infrequently in a corpus, we also can extract it by matching their senses. That is, ontology can compensate the loss of the statistical methods in terminology extraction.

3.2. Similarity of words within ontology

In order to guage the semantic dependency of individual words in a string sequence, similarity measures within framework of

ontology was employed. Although many methods are proposed to measure semantic similarity between words (Rada, 1989; Richardson et al., 1994; Li, 2003), here a traditional method and a newly developed method are attempted in our study. That is, Rada et al.'s method (Rada, 1989) (referred as Rada method hereafter) and Li et al.'s method (Li, 2003) (referred as Li method hereafter).

In Rada method, similarity of two words is measured by the length of the shortest path between them in the hierarchical tree. The basic idea behind this method is very intuitively: words are associated with concepts in the "is a" (ISA) hierarchy, therefore, we can find the first concept in the hierarchical semantic network that subsumes the concepts containing the compared words and then a path that can connect these two words is found. For example, we can see from Fig. 1 that "person" is the mutual subsumer of "boy" and "girl", "teacher" and "girl" in the semantic hierarchy. However, the shortest path between "boy" and "girl" is "boy"–"male"–"person"–"female"–"girl" and the length of this path is four, while the shortest path between "girl" and "teacher" is "girl"–"female"–"person"–"adult"–"professional"–"educator"–"teacher" and the length of this path is six, so we can say that "boy" is more similar with "girl" than "teacher". In the case that a word is polysemous (i.e., a word having many meanings), multiple paths may exist between the two words. Only the shortest path is used to calculate semantic similarity between them.

In this paper, the similarity between words using Rada method is calculated as with the following formula.

$$\mathrm{sim}(w_1, w_2) = e^{-\alpha l} \tag{1}$$

where α is a predefined constant and l is the length of the shortest path of word w_1 and w_2 in the hierarchical tree. The exponential form in similarity calculation is adopted because of Shepard's law which claims that exponential-decay functions are a universal law of stimulus generalization for psychological science (Li, 2003).

The difference of Li method from Rada method is in that not only the shortest path between compared words, but also the depth of their subsumer in the ontology hierarchy, and the subsumer's local semantic density are considered to calculate the similarity in Li method. The basic idea behind this method is to overcome the weaknesses in Rada method. For example, if we want to calculate the similarity between "animal" and "girl", we will find the shortest path is "girl"–"female"–"person"–"life form"–"animal", and, the length of this path is the same as "boy" to "girl" as four. This makes Rada method not convincing for measuring the words' similarity. So we should take the depth of words' subsumer in to account. Moreover, local density of the concepts will affect the similarity of two words. For example, if there are few concepts as "life form" occurring in the context of "girl" and "animal", it will make them more similar in that special context.

In this paper, the similarity between words using Li method is calculated with formula (2)–(5).

$$\mathrm{sim}(w_1, w_2) = f_1(l) + f_2(d) + f_3(fr) \tag{2}$$

$f_1(l)$, $f_2(d)$ and $f_3(fr)$ are defined as follows.

$$f_1(l) = e^{-\beta l} \tag{3}$$

where β is a predefined constant and $l$is the length of the shortest path of w_1 and w_2 in the hierarchical tree.

$$f_2(d) = \frac{e^{cd} - e^{-cd}}{e^{cd} + e^{-cd}} \tag{4}$$

where c is a predefined constant and$d$is the depth of the subsumer of w_1 and w_2 in the hierarchical tree. For instance, the depth of "person" as the subsumer of "boy" and "girl" is three in Fig. 1.

$$f_3(fr) = e^{-\gamma/fr} \tag{5}$$

where fr is the frequency of the extracted terminology candidate which contains w_1 and w_2 using JK method. We do not use information content method as specified in Yang and Liu (1999) to calculate the local semantic density due to the sparseness of a specific concept in a document. In addition, the co-occurrence of w_1 and w_2 in documents is considered as an important factor for their being a terminology.

It should be pointed out here that we discussed words' similarity measures within ontology because we want to use them to rank the words' semantic dependency, i.e., the intension of words' semantic correlation. Usually, similarity of a word pair is a value between 0 and 1 with many strict constraints. However, for semantic dependency ranking, what we care about of word pairs is merely the rank of the similarity numbers of the word pairs rather than the real number of similarities. Thus,α could be predefined arbitrarily and the real number of c, β, γ is less important than their relative number.

3.3. The proposed approach for terminology extraction using ontology

Based on JK method and the similarity of words within ontology, a new approach is proposed to use ontology to improve terminology extraction from documents. Here are the main steps of our approach and the details will be discussed later.

1. Extract the repetitions from documents.
2. Conduct POS (part of speech) processing for repetitions and extracted patterns from repetitions using JK regular expression.
3. If an extracted pattern is a collocation already included in ontology hierarchy such as "professional person", then it will be accepted as a terminology. Otherwise, similarity dependency will be given for this pattern.
4. Accept the patterns whose semantic dependencies are greater than the critical semantic dependency on the point of retaining proportion (RP) as terminologies. RP is a predefined threshold for patterns' proportion with highest semantic dependency at a ratio.

Input:

s_1, the first sentence

s_2, the second sentence

Output:

Multi-word extracted from s_1 and s_2.

Procedure:

```
s1 = {w1,w2,…,wn}, s2 = {w1',w2',…,wm'}, k=0
For each word wi in s1
    For each word wj' in s2
        While(wi equal to wj')
            k++
        End while
        If k>1
            extract the words from wi to wi+k to form a repetition
            k = 0
        End if
    End for
End for
```

Fig. 2. The algorithm used for repetition extraction from sentences.

In Step 1, the repetitions are extracted by matching the same sequences between two sentences. For example, if we have the following two sentences:

- Standard oil co and bp north America inc said they plan to form a venture to manage the money market borrowing and investment activities of both companies.
- The venture will be called bp/standard financial trading and will be operated by standard oil under the oversight of a joint management committee.

From the above two sentences, "*standard oil*" will be extracted as a repetition. The algorithm we used for extracting repetitions from sentences is shown in Fig. 2.

In Step 2, we conduct the POS tagging for repetitions using QTAG which is a probabilistic POS tagger and can be downloaded freely. (online: http://www.english.bham.ac.uk/staff/omason/software/qtag.html).

In Step 3, semantic dependency of a pattern is produced as the maximum similarity, which is measured by either Rada method or Li method, of two nouns from the pattern. That is, assuming p is a pattern, and $(w_1, w_2, \ldots, w_n)$ is the noun words from p, the semantic dependency of p is defined as follows.

$$sd(p) = \max[\mathrm{sim}(w_i, w_j)] \quad i \neq j, 1 \leqslant i, j \leqslant n \tag{6}$$

In Step 4, RP is set by trial and error practice.

4. Experiment and evaluation

In this section, a series of experiments were carried out to evaluate our method on terminology extraction from English documents. Both Rada method and Li method were used to rank the semantic dependencies of the noun words in patterns.

4.1. Corpus

Reuters-21578 text collection (online: http://www.daviddlewis.com/resources/test-collections/reuters21578/) was applied as our experimental data. It appeared as Reuters-22173 in 1991 and was indexed with 135 categories by personnel from Reuters Ltd. in 1996. By our statistics, it contains in total 19403 valid texts with average 5.4 sentences for each text. Because these documents are mostly short passages and there are not enough sentences in each one of them to extract the repetitions, we only fetched out 196 documents whose sizes are larger than 4 K from the corpus.

4.2. Experimental design

For convenience of evaluation, a standard terminology base for 30 documents randomly selected from the target 196 documents is constructed manually. In order to extract repetitions from documents, individual sentences are aligned using the sentence boundary determination method described in Weiss (2004). Thus, 8694 sentences with 139,836 words are aligned for the 196 target documents. Then the repetition extraction method depicted in Fig. 2 was utilized and 7945 repetitions were produced. Next, QTAG is used to conduct the POS tagging for repetitions and the regular expression in JK method employed to extract the patterns, i.e. the final terminologies in JK method, and the evaluation is conducted by comparison with the standard terminology base as shown in Fig. 3.

The overall process using ontology methods to improve terminology extraction was depicted in Fig. 4. The similarity measure of Rada method and Li method were used to rank the similarity dependency of patterns, respectively. All the parameters in formula (1), (3)–(5) are set equal to 1, i.e., $\alpha = \beta = c = \gamma = 1$, because we merely want to rank the semantic dependencies among the noun words in patterns other than to measure the real similarities of them which are should be required to conform to the human psychology. Here, three exceptions should be noted when we used the ontology method for semantic dependency ranking for the patterns. The first one is that there are ANs in the patterns so that no noun pair can be used to calculate the similarity. The second one is that there some nouns unregistered in WordNet so we can not determine their dependencies of other nouns (for example, the trademarks such as "Microsoft", "Nomura", etc). The third one is that there are some nouns between them we can not establish a relationship automatically with WordNet (For example, neither "computer terminal" can be found in collocations of WordNet 1.5 nor "computer" and "terminal" can be linked by its ontology hierarchy). For the patterns with first and second type of exception, we regard them as terminologies directly. For the third case, we can manually select the correct terminologies due to their limited number.

4.3. Evaluation

Tables 1 and 2 are the average precisions and recalls of terminology extraction with ontology support. Compared with the average precision and recall of JK method depicted in Fig. 3, we can see that the average precision is improved with the ontology methods. And Li method can achieve better average precision than Rada

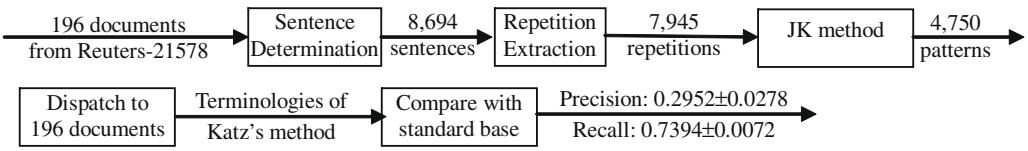

Fig. 3. Terminology extraction by JK method and its performance evaluation.

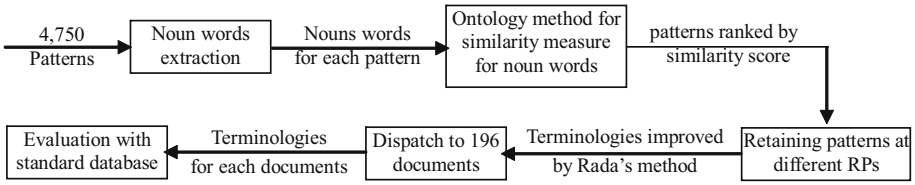

Fig. 4. The process of terminology extraction improved by ontology methods.

Table 1
The precisions of terminology extraction with ontology methods at different RPs.

RP	0.8	0.6	0.4
Rada	0.3155 ± 0.0454	0.3256 ± 0.0440	0.3323 ± 0.0400
Li	0.3274 ± 0.0359	0.3493 ± 0.0406	0.3629 ± 0.0371

Table 2
The recall of terminology extraction with ontology methods at different RPs.

RP	0.8	0.6	0.4
Rada	0.7059 ± 0.0034	0.7033 ± 0.0084	0.7024 ± 0.0110
Li	0.7254 ± 0.0081	0.7191 ± 0.0102	0.7054 ± 0.0119

method. It is reasonable that the recall produced by JK method is the highest one because the terminologies produced by ontology methods are originated from the terminologies produced by JK method.

In addition, we carried out a series of *t*-tests as specified in Yang and Liu (1999) to observe the significance of the performance of each method. Tables 3 and 4 show the results of *t*-tests and the following codification of the *P*-value in ranges was used: "$\gg$" and "$\ll$" mean that the *P*-values is lesser than or equal to 0.01, indicating a strong evidence of that a system generates a greater or smaller extraction error than another one, respectively; "<" and ">" mean that *P*-value is bigger than 0.01 and minor or equal to 0.05, indicating a weak evidence that a system generates a greater or smaller classification error than another one, respectively; "$\sim$" means that the *P*-value is greater than 0.05 indicating that it does not have significant differences in the performances of the systems. It can be seen from Table 3 that on precision, ontology methods can produce significantly higher performance than Katz's method. Moreover, when RP improves, the precision of ontology methods can also be improved. It can also be drawn from Table 4 that Li method can achieve a bit superior recall than Rada method although this point is not very significant.

The improvement in precision with ontology methods demonstrates that terminology exactly has closer semantic relationship among its constituents than non-lexical NP. Moreover, the semantic relationship of words within ontology should be measured by combing various factors such as the ingredients in Li method other than purely the shortest path in hierarchy tree. Furthermore, although ontology methods will cause a loss in recall of terminology extraction as a cost of improving precision, this kind of loss is trivial and worthwhile in practical application when mass documents are confronted with and the recall will be improved automatically by the corpus size.

Table 3
Results of t-test on precision of each method for terminology extraction.

Method	Rada (0.8)	Rada (0.6)	Rada (0.4)	Li (0.8)	Li (0.6)	Li (0.4)
Katz	$\ll$	$\ll$	$\ll$	$\ll$	$\ll$	$\ll$
Rada (0.8)		$\ll$	$\ll$	$\ll$	$\ll$	$\ll$
Rada (0.6)			<	$\sim$	$\ll$	$\ll$
Rada (0.4)				$\sim$	$\ll$	$\ll$
Li (0.8)					$\ll$	$\ll$
Li (0.6)						$\ll$

Table 4
Results of t-test on recall of each method for terminology extraction.

Method	Rada (0.8)	Rada (0.6)	Rada (0.4)	Li (0.8)	Li (0.6)	Li (0.4)
Katz	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$
Rada (0.8)		$\sim$	$\sim$	$\ll$	<	$\sim$
Rada (0.6)			$\sim$	$\ll$	$\ll$	<
Rada (0.4)				$\ll$	$\ll$	$\ll$
Li (0.8)					>	$\gg$
Li (0.6)						>

5. Concluding remarks and future work

In this paper, we adopt ontology to improve the performance of terminology extraction from documents. Firstly, we present a review of current trends on terminology extraction. Secondly, ontology and two popular methods of words similarity measure within the framework of ontology were introduced. Then, JK method was adopted to extract the terminological candidates from documents and ontology methods were adopted to rank the semantic dependencies of the terminological candidates with the goal to improve extraction precision. Finally, we carried out a series experiments on terminology extraction from the Reuters-21578 corpus.

The experimental results demonstrate that: (1) JK method can produce a high recall on terminology extraction as more than 70% on Reuters-21578 corpus; (2) Ontology can improve the precision of terminology extraction significantly at an acceptable cost in recall with all the different parameter settings; (3) Li method for words similarity measure within ontology outperforms Rada method in ranking noun words' similarity dependency.

Although some initial work on using ontology for terminology extraction is attempted in this paper, some problems are still left in our method. The first one is how to keep the recall stable as applying ontology for terminology extraction. Secondly, how to map the semantic dependency of words in terminologies into the words' relationship in the ontology appropriately should be explored. That is, to develop a terminology oriented ranking method with the support of background knowledge. Thirdly, parameters such as c, β, γ and RP should be optimally learned automatically from specific corpus.

As far as future work is concerned, terminology extraction is still of our interest. We will combine the statistical and linguistic methods based on their superiorities for solving this problem. On the other hand, ensemble ontology method such as combing Mesh ontology and WordNet ontology will be attempted for text clustering and information retrieval, so that the contextual and background knowledge can be integrated into practical intelligent information processing applications.

Acknowledgements

This work is supported by Ministry of Education, Culture, Sports, Science and Technology of Japan under the "Kanazawa Region, Ishikawa High-Tech Sensing Cluster of Knowledge-Based Cluster Creation Project" and partially supported by the National Natural Science Foundation of China under Grant No. 70571078 and 70221001.

References

Bloehdorn, S., & Hotho, A. (2004). Boosting for text classification with semantic features. In *Proceedings of the workshop on mining for and from the semantic web at the 10th ACM SIGKDD conference on knowledge discovery and data mining (KDD 2004)* (pp. 70–87).

Chen, K., & Chen, H. (1994). Extracting noun phrases from large-scale texts: A hybrid approach and its automatic evaluation. In *Proceedings of 32nd annual meeting of the association for computational linguistics* (pp. 234–241). Las Cruces, New Mexico.

Choueka, Y. et al. (1983). Automatic retrieval of frequent idiomatic and collocational expressions in a large corpus. *Journal for Literary and Linguistic Computing, 4*, 34–38.

Church, K. W. et al. (1989). Parsing, word association and typical predicate-argument relations. In *Proceedings of the international workshop on parsing technologies* (pp. 103–112). Pittsburgh, PA: Carnegie Mellon University.

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics, 16*(1), 22–29.

Correa, R. F., & Ludermir, T. B. (2006). Improving self-organization of document collections by semantic mapping. *Neurocomputing, 70*, 62–69.

Hotho, A. et al. (2003). Ontologies improve text document clustering. In *Proceedings of the third IEEE international conference on data mining* (pp. 541–544).

Justeson, J. S., & Katz, S. M. (1995). Technical terminology: Some linguistic properties and an algorithm for identification in text. *Natural Language Engineering, 1*(1), 9–27.

Kohler, J. et al. (2006). Ontology based text indexing and querying for the semantic web. *Knowledge-Based Systems, 19*, 744–754.

Li, Y. et al. (2003). An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on Knowledge and Data Engineering, 15*(4), 871–882.

Li, Y. et al. (2008). Text document clustering based on frequent word meaning sequences. *Data and Knowledge Engineering, 64*, 381–404.

Miller, G. A. (1998). Nouns in WordNet. In C. Fellbaum (Ed.), *WordNet: An electronic lexical database* (pp. 23–46). The MIT Press.

Ontology. http://en.wikipedia.org/wiki/Ontology.

Rada, R. et al. (1989). Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man and Cybernetics, 19*(1), 17–30.

Richardson, R. et al. (1994). Using WordNet as a knowledge base for measuring semantic similarity between words. In *Techinical report working paper CA-1294*. Dublin, Ireland: School of Computer Applications, Dublin City University.

Rodriguez, M. B. et al. (2000). Using WordNet to complement training information in text categorization. In *Proceedings of RANLP-97, Proceedings of second international conference on recent advances in natural language processing II: Selected papers from RANLP'97, Current issues in linguistic theory (CILT)* (Vol. 189, pp. 353–364).

Scott, S., & Matwin, S., (1998). Text classification using WordNet hypernyms. In *Proceedings of the COLING/ACL 98 workshop on usage of WordNet, Natural language processing systems* (pp. 45–52).

Smadja, F. (1993). Retrieving collocations from text: Xtract. *Computational Linguistics, 19*(1), 143–177.

Weiss, S. M. et al. (2004). *Text mining: Predictive methods for analyzing unstructured information*. New York: Springer-Verlag (pp. 36–37).

Yang, Y. M., & Liu, X. (1999). A re-examination of text categorization methods. In *Proceedings on the 22nd annual international ACM SIGIR conference on research and development in information retrieval* (pp. 42–49). Berkeley, CA.

Zhou, X. et al. (2007). Topic signature language models for ad hoc retrieval. In *IEEE transactions on knowledge and data engineering* (Vol. 19 (9), pp. 1–12).