Term Extraction Using Verb Co-occurrence

Teruo KOYAMA and Kyo KAGEURA
Human and Social Information Research, National Institute of Informatics
2-1-2, Hitotsubashi, Chiyoda-ku
101-8430, Tokyo
Japan
{t_koyama,kyo}@nii.ac.jp

Abstract

This paper proposes a method of automatic term extraction in which documents are grouped on the basis of discourse/domains by means of characteristic verbs and nominal terms. The method aims to (a) extract terms in accordance with their general positions in the discourse, (b) enhance the precision of extraction and (c) cover relatively low-frequency terms in extraction. Experiments show that it performs well in terms of these objectives.

1 Introduction

Term extraction is important in natural language processing, especially in addressing the lexical bottleneck. To date, various statistical methods, such as TF-IDF, have been applied to this task, and have attained great success (Aizawa, 2003; Bourigault et al., 1998; Daille et al., 1994; Kageura & Koyama, 2000; Mima & Ananiadou, 2000; Moens, 2000; Nakagawa, 2000).

However, some problems still remain: (a) the concept of "domain" or "field" is determined mostly in advance; (b) the positions of extracted terms in the domain are not very clear; (c) generally, in statistical approaches, the results are sensitive to frequency, and it is difficult to deal with low-frequency terms.

This paper proposes a way of dealing with these problems. The basic idea is that there are some characteristic verbs used in scientific discourse, which can be used (a) to define the domains/sub-domains in the corpora of the broad domain and (b) to estimate the positions[1] of terms that co-occur with them using clustering analysis.

Some studies have analyzed the role of verbs in documents (Eumeridou et al., 2002; Klavans & Kan, 1998). Most of them focus on the micro-structure of discourse, such as verb-argument relations. Our approach aims at elucidating the positions of verbs as a deciding factor in the macro discoursal structure[2] of documents in corpora.

The application of clustering in extracting various generic characteristics in language has been intensively carried out (Pereira et al., 1993; Utsuro et al., 1998). In contrast to these works, the present study is specifically concerned with macro-discourse and lexical clues that reflect it.

In the following, section 2 outlines the method we propose. Section 3 will discuss how the combination of major verbs can be used to classify terms and documents; section 4 will discuss the effect of using the proposed method in term extraction.

2 Outline of the proposed method

Our basic hypothesis is:

> For a coherent domain, sub-domain or set of documents, there are a few major "verbal" concepts and "nominal" concepts that determine the macro-structure of discourse, which are respectively represented by characteristic verbs and nouns.

Based on this hypothesis, we define the procedure of identifying macro-discourse and extracting terms as follows:

1. Extract verbs [3] and term candidates (noun related sequences) which are expected to represent macro-discourse.
2. Cluster the term candidates based on the co-occurring verbs extracted in step 1.

[1] Here, by "position", we mean to which discoursal class, sub-domain or conceptual group a term belongs. See Sowa (1993) for some related discussion about the position of concepts represented in terms.

[2] By "macro" we mean the generalized discoursal structure that characterizes a domain, a sub-domain or a group of documents.

[3] Here, we include verb stems as well according to Japanese word classification.

3. Manually identify groups of term candidates that correspond to macro-discourse groups, and identify representative vectors consisting of verbs that correspond to discoursal groups on the basis of step 2.
4. Cluster documents on the basis of verb vectors defined in step 3.
5. For each clustered document, apply a basic term extraction method.

By identifying discourse groups within given corpora, it is possible to macro-structurize terms. Also, quantitative information is expected to become relatively more sensitive to low-frequency term candidates.

In the following experiments, we use a corpus extracted from the database edited by NII (National Institute of Informatics), consisting of research reports in various scientific fields supported by grant-in-aid from the Japan Society for the Promotion of Science for 1998. We use two sub-corpora, i.e. in the field of engineering and in natural science.

The number of reports is 7753 for engineering and 5171 for natural science. Each report consists of about 750 (Japanese) characters.

We applied the "Juman" morphological analyzer and extracted sequences of nouns or word stems as (noun) term candidates.

3 Defining contextual information

Here, we elaborate on steps 1 to 3 described in section 2.

Step#1.

First, we order candidate terms and verbs by TF-IDF, in order to select important terms and verbs in the domain. We select the 15 verbs and 50 candidate terms with the highest TF-IDF values.

Step#2.

In relation to a verb v, we can select a document set D_v which includes the verb. For this set, a parameter that indicates the strength of the co-occurrence tendency between a term t and v can be defined as

$$R_{tv} = \frac{\frac{f_v}{d_v}}{\frac{f_h}{d_h}} - 1$$

Where, [4]

- f_v: term frequency in D_v
- f_h: term frequency in all documents (D_h)
- d_v: document number of D_v
- d_h: whole document number of D_h

Step#3.

Using multiple verbs, we can obtain a vector $\overrightarrow{R_t} = \{R_{tv}\}$ for the term t. Along we select 15 verbs, the 50 terms distribute themselves in a 15-dimensional vector space. On these vectors, we apply clustering analysis using the Euclidean distance and compact clustering method. Figure 1 shows the result of the analysis for the engineering corpus.

Cutting the tree at level 3, in Fig.1, we obtain 8 term groups. Each group can be presumed to be chemistry-1 (group1), material engineering (group2), civil engineering (group3), semiconductors (group4), chemistry-2 (group5), imaging (group6), system engineering/control (group7), and condensed matter physics (group8) respectively. This grouping seems to correspond to real engineering subdomains.

Applying the same method to the natural science corpus, we obtain 6 term groups, namely, chemistry-1, astronomy/geology, biology, mathematics, physics, and chemistry-2.

4 Term extraction experiment

Once the grouping of terms has been done, we can obtain mean (term) vectors for each group. These vectors can be used to evaluate the strength of the relation between a document and each term group. Suppose a vector $\overrightarrow{V_i} = \{V_{ik}\}$ is the mean vector of the i-th term group, and V_{ik} is the component corresponding to the k-th verb. The strength of the relation between a document D_j and the i-th term group is given by

$$R_{ij} = \prod_{k=1}^{n} M_{ijk} - 1$$

where,

$$M_{ijk} = \left\{ \begin{array}{ll} V_{ik} + 1 & (V_k \in T(D_j)) \\ 1 & (V_k \notin T(D_j)) \end{array} \right\}$$

$T(D_j)$ denotes the set of words in the document D_j.

Using this measure, we can apply steps 4 and 5 in section 2.

Step#4.

[4] This parameter can be regarded to be proportional to the likelihood ratio of the appearance of t within D_v and D_h, minus one.

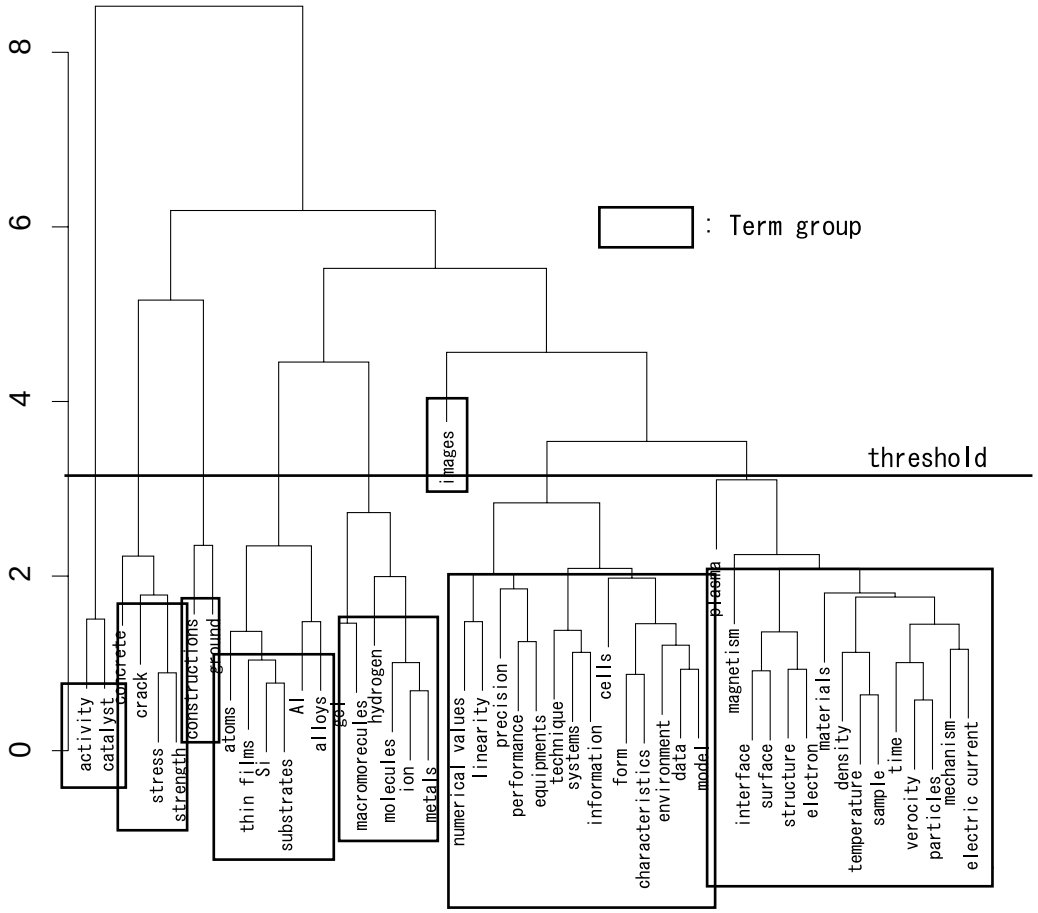

Figure 1: Result of the cluster analysis (engineering)

If we set a proper threshold, we can divide the documents into two groups, namely, documents with a (strong) relation to the term group, and others. For the engineering data, we adopt a threshold value of 0.5.

Step#5.

Once the corpus is divided into two document sets for each discoursal group, we can extract the candidates of terms, based on the division. We tried two term extraction methods. One is for rather frequent terms, and the other is for less frequent terms.

If a (frequent) candidate term has the tendency to appear more frequently in the selected document set, it can be regarded as a term belonging to the corresponding group. To evaluate the tendency, we adopt a ranking score for a candidate term C in association with term group i as

$$S_{ci} = \sqrt{f_i} \times (\frac{\frac{f_i}{d_i}}{\frac{f_h}{d_h}})$$

Where, [5]

- f_i is the candidate term frequency in the document group i (D_i)
- f_h is the candidate term frequency in all the documents (D_h)
- d_i is the document number of D_i
- d_h is the whole document number of D_h

Less frequent terms are extracted according to the following criterion: if (i) the total frequency of a candidate is greater than or equal to 5, and (ii) over 90% of the tokens appear in the document set, then it is considered a term.

Using these methods, we can extract terms corresponding to the discoursal domain represented by term groups.

To evaluate the precision of the methods, we count the number of non-term candidates in-

[5]This value is defined as the likelihood ratio modified by the candidate frequency within the selected document group. The modification is carried out to avoid the noise caused by low-frequency candidates.

Table 1: Number of non-terms among 100 candidates (in engineering)

	documents	freq.	less-freq.
Group1	1818	9	11
Group2	2596	4	12
Group3	2078	7	8
Group4	2485	8	3
Group5	1616	10	9
Group6	1393	7	4
Group7	1326	8	8
Group8	1363	8	11
TF-IDF	-	27	
First 3800	3800	31	

cluded in the top 100 results for frequent terms, and in randomly sampled 100 results for less frequent terms.

We also prepare data on the top 100 TF-IDF value and the result of the proposed method applied to the first 3800 documents in the corpus. Table 1 shows the result.

It shows that the precision of term extraction based on document division is better than using the TF-IDF score and random grouping of documents. We applied the same method to 6 groups in the natural science field, and observed a similar improvement.

5 Conclusion and discussion

We have proposed a method of improving simple automatic term extraction based on grouping documents by discourse/domains using co-occurrences of characteristic verbs and nominal terms.

The underlying hypothesis was: If we focus on a few major "verbal" concepts and "nominal" concepts that are expected to determine the macro-structure of discourse, which are respectively represented by characteristic verbs and nouns, we could structurize or group the documents in the corpus according to the macro-structure of discourse or topic, which in turn will contribute to enhancing term extraction.

The experimental results based on the corpora, of engineering and of natural science have shown that the method is highly promising. We are currently applying the method to the corpus of human science. Though a detailed evaluation has not yet been carried out, the results so far are highly promising.

The current method still relies on some intuition and manual evaluation in the intermediate stages. We are examining ways of automatizing these phases as well as improving the methods of the term extraction phase using different weighting measures.

References

Akiko N. Aizawa. 2003. "An information-theoretic perspective of tf-idf measures," *Information Processing and Management*, 39(1): 45–65.

Bourigault, D., Jacquemin, C., and L'Homme, M. C. 1998. *Proceedings of the 1st International Workshop on Computational Terminology (Computerm 1998)*. COLING-ACL, Montreal.

Daille, B., Gaussier, E. and Langé, M. 1994. "Towards automatic extraction of monolingual and bilingual terminology," *COLING-94*: 515–521.

Eumeridou, E., Nkwenti-Azeh, B. and McNaught, J. 2002. "The contribution of verbal semantic content towards recognition," *International Journal of Corpus Linguistics*, 7(1): 87–106.

Kageura, K. and Koyama, T. eds. 2000. *Special Issue on Japanese Term Extraction: Terminology*, 6(2).

Klavans, J. and Kan, M. Y. 1998. "Role of verbs in document analysis," *COLING-ACL'98*: 680–686.

Mima, H. and Ananiadou, S. 2000. "An application and evaluation of the C/NC-value approach for the automatic term recognition of multi-word units in Japanese," *Terminology*, 6(2): 175–194.

Moens, M. F. 2000. *Automatic Indexing and Abstracting of Document Texts*. Dordrecht: Kluwer.

Nakagawa, H. 2000. "Automatic term recognition on statistics of compound nouns," *Terminology*, 6(2): 175–194.

Pereira, F., Tishby, N., and Lee, L. 1993. "Distributional clustering of English words," *ACL-92*: 128–135.

Sowa, J. F. 1993. "Lexical structures and conceptual structures." In Pustejovsky, J. ed. *Semantics and Lexicon*. Dordrecht: Kluwer. 223–262.

Utsuro, T., Miyata, T., and Matsumoto, Y. 1998. "General-to-specific model selection for subcategorization preference," *COLING-ACL'98*: 1314–1320.