Collecting Novel Technical Terms from the Web by Estimating Domain Specificity of a Term

Takehito Utsuro[1], Mitsuhiro Kida[2], Masatsugu Tonoike[3], and Satoshi Sato[4]

[1] Graduate School of Systems and Information Engineering, University of Tsukuba, 1-1-1, Tennodai, Tsukuba, 305-8573, Japan
[2] Nintendo Co., Ltd., 11-1, Hokotate-cho, Kamitoba, Minami-ku, Kyoto-shi, 601-8116 Japan
[3] Graduate School of Informatics, Kyoto University, Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, Japan
[4] Graduate School of Engineering, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan

Abstract. This paper proposes a method of domain specificity estimation of technical terms using the Web. In the proposed method, it is assumed that, for a certain technical domain, a list of known technical terms of the domain is given. Technical documents of the domain are collected through the Web search engine, which are then used for generating a vector space model for the domain. The domain specificity of a target term is estimated according to the distribution of the domain of the sample pages of the target term. We apply this technique of estimating domain specificity of a term to the task of discovering novel technical terms that are not included in any of existing lexicons of technical terms of the domain. Out of randomly selected 1,000 candidates of technical terms per a domain, we discovered about 100 $\sim$ 200 novel technical terms.

1 Introduction

Lexicons of technical terms are one of the most important language resources both for human use and for computational research areas such as information retrieval and natural language processing. Among various research issues regarding technical terms, full-/semi-automatic compilation of technical term lexicon is one of the central issues. In various research fields, novel technologies are invented every year, and related research areas around such novel technologies keep growing. Along with such invention of technologies, novel technical terms are created year by year. Considering such a situation, it requires a huge cost for manually compiling lexicons of technical terms for hundreds of thousands of technical domains. Therefore, it is inevitable to invent a technique of full-/semi-automatic compilation of technical term lexicons for various technical domains.

The whole task of compiling a technical term lexicon can be roughly decomposed into two sub-processes: (1) collecting candidates of technical terms of a technical domain, and, (2) judging whether each candidate is actually a technical term of the target technical domain. The technique of the first sub-process is closely related to research on automatic term recognition, and has been relatively well studied so far (e.g., [5]). On the other hand, the technique of the second

Y. Matsumoto et al. (Eds.): ICCPOL 2006, LNAI 4285, pp. 173–180, 2006.

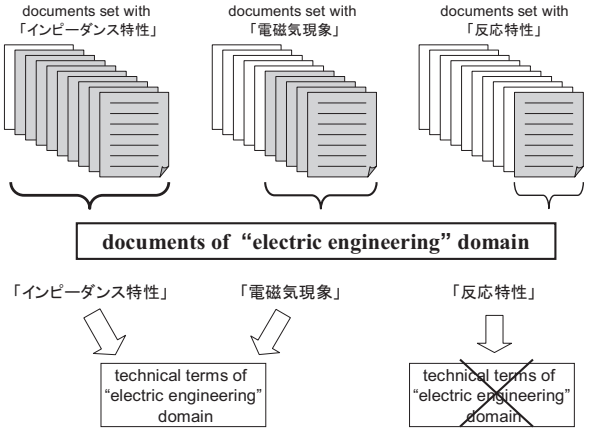

Fig. 1. Degree of Specificity of a Term based on the Domain of the Documents (Example terms: *impedance characteristic*, *electromagnetism*, and *response characteristic*)

sub-process has not been studied well so far. Exceptional cases are works such as [1,2], where their techniques are mainly based on the tendency of technical terms appearing in technical documents of limited domains rather than in documents of daily use such as newspaper and magazine articles. Although the underlying idea of those previous works is very interesting, those works are quite limited in that they require existence of certain amount of technical domain corpus. It is not practical for manually collecting technical domain corpus for hundreds of thousands of technical domains. Therefore, as for the second sub-process here, it is very important to invent a technique for automatically classifying the domain of a technical term.

Based on this observation, among several key issues regarding the second sub-process above, this paper mainly focuses on the issue of estimating the domain specificity of a term. In this paper, supposing that a target technical term and a technical domain are given, we propose a technique of automatically estimating the specificity of the target term with respect to the target domain. Here, the domain specificity of the term is judged among the following three levels: i) the term mostly appears in the target domain, ii) the term generally appears in the target domain as well as in other domains, iii) the term generally does not appear in the target domain.

The key idea of the proposed technique is as follows. In the proposed technique, we assume that sample technical terms of the target domain are available. Using such sample terms with search engine queries, we first collect a corpus of the target domain from the Web. In a similar way, we also collect sample pages that include the target term from the Web. Then, the similarities of the contents of the documents are measured between the corpus of the target domain and each of the sample pages that include the target term. Finally, the domain specificity of the target term is estimated according to the distribution of the domain of those sample pages.

Figure 1 illustrates rough idea of this technique. Among the three example (Japanese) terms, the first term (*impedance characteristic*) mostly appears in the documents of the *"electric engineering"* domain on the Web. In the case of the second term (*electromagnetism*), about half of sample pages collected from the Web can be regarded as in the *"electric engineering"* domain, while the rest are not. On the other hand, in the case of the last term (*response characteristic*), only a few of the sample pages can be regarded as in the *"electric engineering"* domain. In our technique, such difference of the distribution can be easily identified, and the domain specificities of those three terms are estimated.

As experimental evaluation, we first evaluate the proposed technique of estimating domain specificity of a term using manually constructed development and evaluation term sets, where we achieved mostly 90% precision/recall (details are presented in [6]). Furthermore, in this paper, we present the result of applying this technique of estimating domain specificity of a term to the task of discovering novel technical terms that are not included in any of existing lexicons of technical terms of the domain. Candidates of technical terms are first collected from the Web corpus of the target domain. Then, about 70∼80 % of those candidates are excluded by roughly judging the domain of their constituent words. Finally, out of randomly selected 1,000 candidates of technical terms per a domain, we discovered about 100 $\sim$ 200 novel technical terms that are not included in any of existing lexicons of the domain, where we achieved about 75% precision and 80% recall.

2 Domain Specificity Estimation of Technical Terms Using Documents Collected from the Web

In this section, we first describe the proposed technique of estimating domain specificity of a term using the Web.

2.1 Outline

Here, we estimate the domain specificity of a term t with respect to a domain C, supposing that the term t and the domain C are given. Generally speaking, the coarsest-grained classification of domain specificity of a term is binary classification, namely, the class of terms that are used in a certain technical domain, vs. the class of terms that are *not* used in a certain technical domain. In this paper, we further classify the degree $g(t, C)$ of the domain specificity into the following three levels:

$$g(t,C) = \begin{cases} + \ (t \text{ mostly appears in the documents of the domain } C.) \\ \pm \ (t \text{ generally appears in the documents of the domain } C \text{ as well as} \\ \quad \text{in those of the domains other than } C.) \\ - \ (t \text{ generally does not appear in the documents of the domain } C.) \end{cases}$$

(When we simply classify domain specificity of a term into two classes with the coarsest-grained binary classification above, we regard those with domain

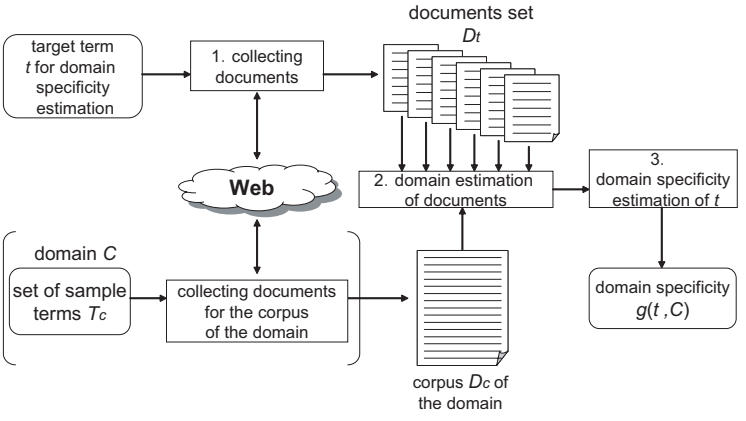

Fig. 2. Domain Specificity Estimation of Terms based on Web documents

specificity '+' or '±' as those that are used in the domain, and those with domain specificity '−' as those that are *not* used in the domain.)

The input and output of the process of domain specificity estimation of a term t with respect to the domain C are given below:

input	target term t for domain specificity estimation, set T_C of sample terms of the domain C
output	domain specificity $g(t, C)$ of t with respect to C

The process of domain specificity estimation of a term is illustrated in Figure 2, where the whole process can be decomposed into two sub-processes: (a) that of constructing the corpus D_C of the domain C, and (b) that of estimating the specificity of a term t with respect to the domain C. In the process of domain specificity estimation, the domain of documents including the target term t is estimated, and the domain specificity of t is judged according to the distribution of the domains of the documents including t. The details of those two sub-processes are described in the followings.

2.2 Constructing the Corpus of the Domain

When constructing the corpus D_C of the domain C using the set T_C of sample terms of the domain C, first, for each term s in the set T_C, we collect into a set D_s the top 100 pages obtained from search engine queries that include the term s.[1] The search engine queries here are designed so that documents that describe the technical term s are ranked high. When constructing a corpus of the Japanese language, the search engine "goo"[2] is used. The specific queries that are used in

[1] Related techniques for automatically constructing the corpus of the domain using the sample terms of the domain include those presented in [4,3]. We are planning to evaluate the performance of those related techniques and compare them with the one employed in this paper.

[2] http://www.goo.ne.jp/

this search engine are phrases with topic-marking postpositional particles such as "*s*-toha," "*s*-toiu," "*s*-wa," and an adnominal phrase "*s*-no," and "*s*."

Then, union of the sets D_s for each s is constructed and denoted as $D(T_C)$:

$$D(T_C) = \bigcup_{s \in T_C} D_s$$

Finally, in order to exclude noise texts from the set $D(T_C)$, the documents in the set $D(T_C)$ are ranked according to the number of sample terms (of the set T_C) that are included in each document. Through a preliminary experiment, we decided here that it is enough to keep top 500 documents, and regard them as the corpus D_C of the domain C.[3]

2.3 Domain Specificity Estimation of Technical Terms

Given the corpus D_C of the domain C, domain specificity of a term t with respect to a domain C is estimated through the following three steps:

Step 1. Collecting documents that include the term t from the Web, and constructing the set D_t of those documents.

Step 2. For each document in the set D_t, estimating its domain by measuring similarity against the corpus D_C of the domain C. Then, given a certain lower bound L of document similarity, from D_t, extracting documents with large enough similarity values into a set $D_t(C, L)$.

Step 3. Estimating the domain specificity $g(t, C)$ of t using the document set $D_t(C, L)$ constructed in the step 2.

Details of those three steps are given below:

Collecting Web Documents Including the Target Term. For each target term t, documents that include t are collected from the Web. According to a procedure that is similar to that of constructing the corpus of the domain C described in section 2.2, the top 100 pages obtained with search engine queries are collected into a set D_t.

Domain Estimation of Documents. For each document in the set D_t, its domain is estimated by measuring similarity against the corpus D_C of the domain C. Then, given a certain lower bound L of document similarity, documents with large enough similarity values are extracted from D_t into the set $D_t(C, L)$ [6].

Domain Specificity Estimation of a Term. The domain specificity of the term t with respect to the domain C is estimated using the document sets D_t

[3] In our evaluation, about 80∼90 % of the documents of D_C are actually those of the domain C. Even with D_C having all of its documents as of the domain C, we achieved almost the same performance of domain specificity estimation of a term.

and $D_t(C, L)$. Here, this is done by simply calculating the following ratio r_L of the numbers of the documents within the two sets:

$$r_L = \frac{|D_t(C, L)|}{|D_t|}$$

Then, by introducing the two thresholds $a(\pm)$ and $a(+)$ for the ratio r_L, the specificity $g(t, C)$ of t is estimated with the following three levels:

$$g(t, C) = \begin{cases} + \ (a(+) \leq r_L) \\ \pm \ (a(\pm) \leq r_L < a(+)) \\ - \ (r_L < a(\pm)) \end{cases}$$

In experimental evaluation of section 4, as in the case of the lower bound L of the document similarity, the two thresholds $a(\pm)$ and $a(+)$ are also determined using the development term set mentioned above.

3 Collecting Novel Technical Terms of a Domain from the Web

This section illustrates how to apply the technique of domain specificity estimation of technical terms to the task of discovering novel technical terms that are not included in any of existing lexicons of technical terms of the domain. First, as shown in Figure 3, from the corpus D_C of the domain C, candidates of technical terms are collected. In the case of the Japanese language, as candidates of novel technical terms, we collect compound nouns with frequency counts five or more,

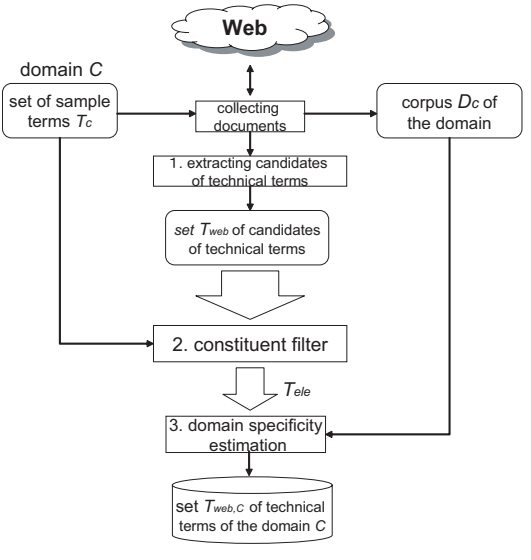

Fig. 3. Collecting Novel Technical Terms of a Domain from the Web

consisting of more than one noun. Here, we collect compound nouns which are not included in any of existing lexicons of technical terms of the domain. Then, after excluding terms which do not share constituent nouns against the sample terms of the given set T_C, the domain specificity of the remaining terms are automatically estimated. Finally, we regard terms with domain specificity '+' or '±' as those that are used in the domain, and collect them into the set $T_{web,C}$.

4 Experimental Evaluation

We evaluate the proposed method with five sample domains, namely, *"electric engineering", "optics", "aerospace engineering", "nucleonics"*, and *"astronomy"*. For each domain C of those five domains, the set T_C of sample (Japanese) terms is constructed by randomly selecting 100 terms[4] from an existing (Japanese) lexicon of technical terms for human use. We evaluate the results of discovering novel technical terms that are not included in any of existing lexicons of technical terms of the domain. First, Table 1 compares the numbers of candidates of novel technical terms collected from the Web, with those after excluding terms which do not share constituent nouns against the sample terms of the given set T_C. As shown in the table, about 70∼80 % of the candidates are excluded, while the rate of technical terms within the remaining candidates increased. This result clearly shows the effectiveness of the constituent noun filtering technique in reducing the computational time of discovering fixed number of novel technical terms. Then, per a domain, we randomly select 1,000 of those remaining candidates, and estimate their domain specificity by the proposed method. After manually judging the domain specificity of those 1,000 terms, we measure the precision/recall of the proposed method as in Table 2, where we achieved about 75% precision and 80% recall. Here, however, as candidates of technical terms, we simply collect compound nouns, where sometimes their term unit is not correct since the technical term candidate could be with a certain prefix or suffix. Considering this fact, Table 2 also gives the term unit correct rate for those with domain specificity '+' or '±'. Finally, taking this term unit correct rate into account, we can

Table 1. Changes in Number of Technical Term Candidates with Constituent Filter

	before filtering		after filtering	
	# of candidates	# of tech. terms (estimated) (%)	# of candidates	# of tech. terms (estimated) (%)
electric engineering	24,460	1,272 (5.2)	6,623	848 (12.8)
optics	29,090	1,047 (3.6)	6,985	866 (12.4)
aerospace engineering	41,279	660 (1.6)	6,364	458 (7.2)
nucleonics	40,439	890 (2.2)	10,834	650 (6.0)
astronomy	29,240	1,170 (4.0)	5,491	659 (12.0)

[4] Through a preliminary experiment, we conclude that it is not necessary to start with the set T_C of sample terms which has more than 100 sample terms. The number of minimum requirement for the size of T_C varies according to domains.

Table 2. Precision/recall of Collecting Novel Technical Terms

(a) with threshold $a(\pm)$

	precision	recall	term unit correct rate
electric engineering	0.754(399/529)	0.828(399/482)	0.393(157/399)
optics	0.766(454/593)	0.875(454/519)	0.368(167/454)
aerospace engineering	0.797(408/512)	0.739(408/552)	0.402(164/408)
nucleonics	0.685(470/686)	0.953(470/493)	0.377(177/470)
astronomy	0.747(480/643)	0.945(480/508)	0.475(228/480)

(b) with threshold $a(+)$

	precision	recall	term unit correct rate
electric engineering	0.697(168/241)	0.853(168/197)	0.494(83/168)
optics	0.743(234/315)	0.932(234/251)	0.453(106/234)
aerospace engineering	0.666(277/416)	0.936(277/296)	0.502(139/277)
nucleonics	0.580(362/624)	0.981(362/369)	0.406(147/362)
astronomy	0.763(350/459)	0.888(350/394)	0.520(182/350)

conclude that, out of the 1,000 candidates, we discovered about $100 \sim 200$ novel technical terms that are not included in any of existing lexicons of the domain. This result clearly supports the effectiveness of the proposed technique for the purpose of full-/semi-automatic compilation of technical term lexicons.

5 Concluding Remarks

This paper proposed a method of domain specificity estimation of technical terms using the Web. We then applied this technique of estimating domain specificity of a term to the task of discovering novel technical terms that are not included in any of existing lexicons of technical terms of the domain.

References

1. T. M. Chung. A corpus comparison approach for terminology extraction. *Terminology*, 9(2):221–246, 2004.
2. P. Drouin. Term extraction using non-technical corpora as a point of leverage. *Terminology*, 9(1):99–117, 2003.
3. C.-C. Huang, K.-M. Lin, and L.-F. Chien. Automatic training corpora acquisition through Web mining. In *Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence*, pages 193–199, 2005.
4. B. Liu, X. Li, W. S. Lee, and P. S. Yu. Text classification by labeling words. In *Proceedings of the 19th AAAI*, pages 425–430, 2004.
5. H. Nakagawa and T. Mori. Automatic term recognition based on statistics of compound nouns and their components. *Terminology*, 9(2):201–219, 2003.
6. T. Utsuro, M. Kida, M. Tonoike, and S. Sato. Towards automatic domain classification of technical terms: Estimating domain specificity of a term using the Web. In *AIRS 2006*, LNCS: Vol. 4182, pages 633–641. Springer, 2006.