Domain Dependence of Lexical Translation: A Case Study of Patent Abstracts

Hiroyuki Kaji
Department of Computer Science, Shizuoka University
3-5-1 Johoku, Hamamatsu-shi 432-8011, Japan
kaji@inf.shizuoka.ac.jp

Abstract

The domain dependence of translations of nouns in English-to-Japanese patent translation is examined using an automatic method for identifying major translations from a pair of language corpora in the same domain. The method calculates the ratio of the number of associated words of a target word that suggest each translation of the target word to the total number of associated words. This ratio indicates how major a translation is in a domain. Application of the method to a bilingual patent-abstract corpus indicates the necessity and effectiveness of dividing the patent domain into subdomains and adapting a bilingual dictionary to subdomains.

1 Introduction

It is well known that dominant or major translations for a word vary with domains, and bilingual-dictionary adaptation to domains is an effective way to improve the performance of machine translation systems. However, bilingual dictionaries have commonly been adapted to domains on the basis of lexicographers' intuition, which results in a high cost and lack of completeness. To overcome this problem, we have developed a method using bilingual comparable corpora to identify major translations in a domain automatically (Kaji, 2004; Kaji, 2005). The essence of the method is to rank translations of a target word according to the ratio of the number of associated words that suggest each translation to the total number of associated words.

In this paper, we use the above method to examine the domain dependence of translations of nouns in English-to-Japanese patent translation. First, a bilingual patent-abstract corpus is divided into several subcorpora, each of which corresponds to a technology area, and major translations of target words are extracted from each of the subcorpora as well as from the whole corpus. Next, for the whole corpus and each of the subcorpora, the distribution of the ratio of associated words suggesting the most major translation of a target word is shown as well as the distribution of the number of extracted translations per target word. Thus, the necessity and effectiveness of dividing the patent domain into subdomains and adapting a bilingual dictionary to subdomains are demonstrated.

2 Method for Identifying Major Translations in a Domain

2.1 Outline

Our method is based on the assumption that translations of associated words are also associated (Rapp, 1995). The alignment of word associations across languages can reveal which associated word of a target word suggests which of its translations. For example, the alignment of an English word association (*plant*, *culture*) with its Japanese counterpart (植物<*SHOKUBUTSU*>, 栽培<*SAIBAI*>) reveals that, for the target word "*plant*," an associated word "*culture*" suggests the translation "植物<*SHOKUBUTSU*>." Naive word-association alignment methods, however, are not effective in the case of using non-parallel bilingual corpora. They suffer from failure in alignment due to topical-coverage disparity between the corpora of two languages as well as ambiguity in alignment. To overcome these difficulties, our method defines the correlation between a translation and an associated word by using the correlations between the translation and other associated words.

The method consists of the following steps (as shown in Fig. 1). First, word associations are extracted from a corpus of each language by setting a threshold for mutual information between words. Second, pairwise correlation between the second-language translations of a first-language target word and its first-language associated words is calculated iteratively. Third, each associated word is assigned to the translation having the highest correlation with it, and the ratio of the number of associated words assigned to each translation to the total number of associated words (in other words, the ratio of associated words suggesting each translation) is calculated. Finally, translations of the target word are ranked in descending order of the ratio of associated words.

2.2 Extraction of word associations

The mutual information $MI(x, x')$ of a pair of words x and x' is defined by the following formula:

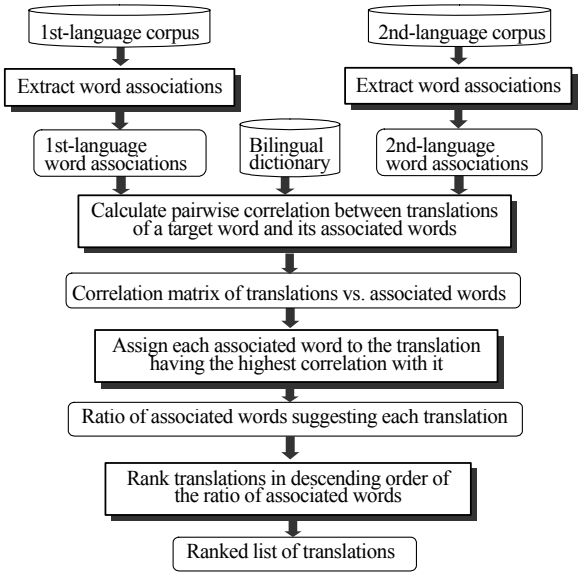


Figure 1. Method for identifying major translations in a domain.

$$MI(x,x') = log \frac{Pr(x,x')}{Pr(x) \cdot Pr(x')},$$

where $Pr(x)$ is the occurrence probability of x, and $Pr(x, x')$ is the co-occurrence probability of x and x' (Church and Hanks, 1990). The occurrence and co-occurrence probabilities are estimated by counting occurrence and co-occurrence frequencies in a corpus. A medium-sized window is used for counting co-occurrence frequencies; in the experiments described in Section 3, the window covered 12 content words on each side of the target word.

A word association is a pair of words having mutual information larger than a threshold θ. Every pair of words (x, x') such that $MI(x, x')>\theta$ is extracted from a corpus of each language. In the experiment described in Section 3, the threshold θ was set at zero.

2.3 Calculation of correlations between translations and associated words

Correlations between translations of a target word and its associated words are defined recursively. A detailed description is given in (Kaji, 2004; Kaji, 2005). The following is an illustrative example. It is assumed that the target word "*plant*" has a set of translations {設備<*SETSUBI*>, 植物<*SHOKUBUTSU*>, プラント<*PURANTO*>} and a set of associated words {*activity*, *bacteria*, *boiler*, *coal*, *computer*, *control*, *culture*, *environment*, *failure*, *flower*, …}. Here, focus was put on the associated word "*culture*."

The correlation C between each of the translations of "*plant*" and the associated word "*culture*" is defined as follows:

- $C(\text{設備} < SETSUBI >, culture) = MI(plant, culture) \cdot \dfrac{PL(\text{設備}, culture)}{\max_{y \in \{\text{設備}, \text{植物}, \text{プラント}\}} PL(y, culture)}$
- $C(\text{植物} < SHOKUBUTSU >, culture) = MI(plant, culture) \cdot \dfrac{PL(\text{植物}, culture)}{\max_{y \in \{\text{設備}, \text{植物}, \text{プラント}\}} PL(y, culture)}$
- $C(\text{プラント} < PURANTO >, culture) = MI(plant, culture) \cdot \dfrac{PL(\text{プラント}, culture)}{\max_{y \in \{\text{設備}, \text{植物}, \text{プラント}\}} PL(y, culture)}$

Furthermore, the plausibility PL that the associated word "*culture*" suggests each of the translations of "*plant*" is defined as follows:

- *PL*(設備<*SETSUBI*>, *culture*)
  = *w*(設備, *plant*, *culture*, *activity*) · *C*(設備, *activity*)
  + *w*(設備, *plant*, *culture*, *bacteria*) · *C*(設備, *bacteria*)
  + *w*(設備, *plant*, *culture*, *boiler*) · C(設備, *boiler*)
  + …
  - *w*(設備, *plant*, *culture*, *x*′) = 1+α
    --- *x*′ is associated with both "*plant*" and "*culture*," and moreover, at least one translation of *x*′ is associated with both "設備" and a translation of "*culture*."
  - *w*(設備, *plant*, *culture*, *x*′) = 1
    --- *x*′ is associated with both "*plant*" and "*culture*," but none of the translations of *x*′ are associated with both "設備" and a translation of "*culture*."
  - *w*(設備, *plant*, *culture*, *x*′) = 0
    --- otherwise.
- *PL*(植物<*SHOKUBUTSU*>, *culture*)
  = *w*(植物, *plant*, *culture*, *activity*) · *C*(植物, *activity*)
  + *w*(植物, *plant*, *culture*, *bacteria*) · *C*(植物, *bacteria*)
  + *w*(植物, *plant*, *culture*, *boiler*) · C(植物, *boiler*)
  + …
  - *w*(植物, *plant*, *culture*, *x*′) = 1+α
    --- *x*′ is associated with both "*plant*" and "*culture*," and moreover, at least one translation of *x*′ is associated with both "植物" and a translation of "*culture*."
  - *w*(植物, *plant*, *culture*, *x*′) = 1
    --- *x*′ is associated with both "*plant*" and "*culture*," but none of the translations of *x*′ are associated with both "植物" and a translation of "*culture*."

· w(植物, *plant*, *culture*, x') = 0
--- otherwise.

- PL(プラント<*PURANTO*>, *culture*)
  = w(プラント, *plant*, *culture*, *activity*) · C(プラント, *activity*)
  + w(プラント, *plant*, *culture*, *bacteria*) · C(プラント, *bacteria*)
  + w(プラント, *plant*, *culture*, *boiler*) · C(プラント, *boiler*)
  + …

  · w(プラント, *plant*, *culture*, x') = 1+α
  --- x' is associated with both "*plant*" and "*culture*," and moreover, at least one translation of x' is associated with both "プラント" and a translation of "*culture*."

  · w(プラント, *plant*, *culture*, x') = 1
  --- x' is associated with both "*plant*" and "*culture*," but none of the translations of x' are associated with both "プラント" and a translation of "*culture*."

  · w(プラント, *plant*, *culture*, x') = 0
  --- otherwise.

The correlations are calculated iteratively with initial values:

C(設備<*SETSUBI*>, *culture*) = C(植物<*SHOKUBUTSU*>, *culture*) = C(プラント<*PURANTO*>, *culture*) = MI(*plant*, *culture*),

C(設備, *activity*) = C(植物, *activity*) = C(プラント, *activity*) = MI(*plant*, *activity*),

C(設備, *bacteria*) = C(植物, *bacteria*) = C(プラント, *bacteria*) = MI(*plant*, *bacteria*),

C(設備, *boiler*) = C(植物, *boiler*) = C(プラント, *boiler*) = MI(*plant*, *boiler*), etc.

C(植物<*SHOKUBUTSU*>, *culture*) probably becomes larger than both C(設備<*SETSUBI*>, *culture*) and C(プラント<*PURANTO*>, *culture*), because:

(1) PL(植物<*SHOKUBUTSU*>, *culture*) naturally has a larger number of terms weighted with (1+α) than both PL(設備<*SETSUBI*>, *culture*) and PL(プラント<*PURANTO*>, *culture*), and

(2) Most of C(植物<*SHOKUBUTSU*>, x') weighted with (1+α) or 1 probably become larger than both C(設備<*SETSUBI*>, x') and C(プラント<*PURANTO*>, x').

It has been proved experimentally that the iterative algorithm works stably for a rather wide range of values of parameter α and the correlations converge rapidly (Kaji and Morimoto, 2005). Table 1 is an example correlation matrix of translations versus associated words calculated for the target word "*plant*."

2.4 Calculation of the ratio of associated words suggesting a translation

The correlation matrix of translations versus associated words is converted into a binary matrix by assigning each associated word to the translation having the highest correlation with it. The resulting binary matrix shows which of the translations is most strongly suggested by each associated word, and the ratio of associated words suggesting each translation is calculated from the binary matrix. For example, the correlation matrix shown in Table 1 results in the binary matrix shown in Table 2. This binary matrix shows that "*coal*" suggests "設備<*SETSUBI*>," "*activity*," "*bacteria*," "*culture*," "*environment*," and "*flower*" suggest "植物<*SHOKUBUTSU*>," and "*boiler*," "*computer*," "*control*," and "*failure*" suggest "プラント<*PURANTO*>."

2.5 Features of the method

One of the main features of the method is that it

Table 1. Correlation matrix of translations vs. associated words for target word "*plant*."

	設備 <*SETSUBI*>	植物 <*SHOKUBUTSU*>	プラント <*PURANTO*>
activity	0.032	***2.104***	0.025
bacteria	0.031	***1.977***	0.018
boiler	2.700	0.053	***2.730***
coal	***2.347***	1.697	2.057
computer	0.707	0.019	***0.726***
control	0.509	0.174	***0.621***
culture	0.052	***3.262***	0.123
environment	1.248	***1.317***	0.051
failure	1.220	0.027	***1.426***
flower	0.055	***4.023***	0.038
:	:	:	:

Table 2. Binary matrix of translations vs. associated words for target word "*plant*."

	設備 <*SETSUBI*>	植物 <*SHOKUBUTSU*>	プラント <*PURANTO*>
activity	0	1	0
bacteria	0	1	0
boiler	0	0	1
coal	1	0	0
computer	0	0	1
control	0	0	1
culture	0	1	0
environment	0	1	0
failure	0	0	1
flower	0	1	0
:	:	:	:

is applicable to non-parallel bilingual corpora. Although non-parallel corpora do not provide translation probabilities directly, our method calculates the ratio of associated words suggesting each translation, which can substitute for a translation probability.

Tanaka and Iwasaki (1996) proposed a method for estimating the translation probability from non-parallel bilingual corpora, but they only demonstrated it in a small-scale experiment. Their method optimizes a translation-probability matrix of first-language vocabulary versus second-language vocabulary, incurring a heavy computational load. In contrast, the method described above is computationally feasible, because it decomposes the problem into the calculation of correlation matrices, each consisting of a few dozen translations versus a few hundred associated words.

Although the method presented here is based on the same assumptions as translation-equivalent extraction methods using contextual similarity (Rapp, 1995; Kaji and Aizono, 1996; Fung and McKeown, 1997; Fung and Yee, 1998; Rapp, 1999), it is quite different from them. It merely ranks translations provided by a bilingual dictionary and it is not able to find new pairs of translation equivalents. However, it has the advantage of being able to calculate the "pseudo" translation probability as well as extract clues for selecting translations (Kaji and Morimoto, 2005).

3 Experiments Using a Patent-Abstract Corpus

3.1 Experimental setting

A series of experiments was carried out using a corpus consisting of Japanese patent abstracts and their English translations. This corpus covers all applications made public in 2003. The total number of abstracts is 348,061. The average length of Japanese abstracts and that of English translations are 597 bytes and 508 bytes, respectively. Although the corpus is in fact a parallel corpus, it was treated as a pair consisting of Japanese and English corpora; namely, neither documents nor sentences were aligned across languages.

The corpus was divided into eight subcorpora, corresponding to the areas indicated by the "Section" part of the International Patent Classification (IPC) code as follows. The number of abstracts included in each subcorpus is shown in parentheses.

A: Human necessities (36,438)
B: Performing operations; transporting (61,405)
C: Chemistry; metallurgy (32,989)
D: Textiles; paper (4,304)
E: Fixed constructions (14,691)
F: Mechanical engineering; lighting; heating; weapons; blasting (29,763)
G: Physics (88,859)
H: Electricity (79,549)

Since the experiments focused on English-to-Japanese translation of nouns, a bilingual noun dictionary was compiled by collecting pairs of nouns from the EDR English-to-Japanese and Japanese-to-English dictionaries (EDR, 1990). The resulting dictionary includes many possible pairs of translation equivalents, i.e., 633,000 pairs of 269,000 English nouns and 276,000 Japanese nouns. It is thus suitable for the experiments.

The method described in Section 2 was applied to subcorpora (A, B, … , H) as well as the whole corpus (ALL). English nouns occurring 50 or more times in the whole corpus were selected as target words. If a target word occurred less than 10 times in a subcorpus, it was deleted from the set of target words for that subcorpus. Accordingly, the numbers of target words for different subcorpora did not coincide.

We intentionally ignored translations of a target word that are rarely used in a (sub)domain by setting the threshold for the ratio of associated words suggesting a translation, abbreviated to '*RAW*,' at 2.5%. That is, translations with *RAW* less than 2.5% were deleted from a ranked list of translations produced by the method described in Section 2.

3.2 Ranked lists of translations obtained for sample target words

Table 3 shows results from the experiment using the whole corpus. Translations of the target word "*plant*" are listed together with their *RAW*s and some of the associated words suggesting them. For comparison, Table 3 also lists the translations of the same target word obtained from a pair of English and Japanese financial-news corpora. These results indicate that the method for identifying major translations works properly.

Table 4 compares the results for five sample target words in the cases of using the whole corpus (ALL) and the subcorpora (A, B, … , H). These results demonstrate that major translations vary from subdomain to subdomain, and a target word sometimes has only one major translation in some subdomains. It is therefore desirable to tune a bilingual dictionary to subdomains. Table 4 also indicates that some target words still have two or more major translations in some subdomains; narrower subdomains may be more suitable.

3.3 Number of translations per target word

The list of translations obtained for a target word consists not of all possible translations, but of translations used in a particular domain. Therefore, the length of the list, i.e., the number of translations, roughly indicates the difficulty of translating the target word in the particular domain. The average numbers of translations per target word in the whole corpus and in the subcorpora are as follows:

Table 3. Translations of "*plant*" in patent domain and financial-news domain.

Domain	#	Translation*	*RAW* (%)	Associated words suggesting the translation
Patent	1	植物 (flora)	46.5	acid, acid sequence, action, activity, animal, aroma, atmosphere, bacteria, bottle, buoyancy, etc.
	2	設備 (apparatus, facilities)	26.4	amount, ash, block, boiler, building, chemical, coal, conditioner, condition, construction, etc.
	3	プラント (industrial plant)	21.1	abnormality, alarm, arithmetic, care, cause, communication line, company, computer, control, control data, etc.
Financial news†	1	工場 (factory)	67.7	Alabama, aluminum, Anderson, annual, Argentina, assembly, assembly plant, auto, auto maker, Ball, etc.
	2	設備 (apparatus, facilities)	19.7	Alberta, building, capacity, Carolina, chemical, Chernobyl, coal, compact, demand, efficiency, etc.
	3	プラント (industrial plant)	6.6	agreement, Asia, Chinese, Co., contract, energy, ethylene, Exxon, gas, Ind., etc.
	4	装置 (apparatus, equipment)	3.5	end, engine, fuel, glass, model, shift, tire, wheel
	5	工場労働者 (factory worker)	2.6	cotton, engineer, GM, labor, Texas, union

* English translations other than "*plant*" are given in parentheses.

† A *Wall Street Journal* corpus (July 1994 to Dec. 1995; 189MB) and a *Nihon Keizai Shimbun* corpus (Dec. 1993 to Nov. 1994; 275MB) were used.

Table 4. Translations of sample target words in the whole domain and subdomains.

Target word	Translation*	*RAW*† (%)								
		ALL	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**
administration	管理 (management, control)	***50.7***	7.8	***100***	14.3	-	-	-	***73.9***	***96.2***
	行政 (government)	-	-	-	-	-	-	-	4.3	-
	局 (government, department)	-	-	-	-	-	-	-	-	3.8
	経営 (management of an organization)	3.1	-	-	-	-	-	-	8.4	-
	運営 (operation)	-	-	-	-	-	-	-	9.7	-
	掌 (conducting, management)	-	3.6	-	-	-	-	-	-	-
	投与 (giving medication)	39.7	***88.6***	-	***85.7***	-	-	-	2.7	-
column	柱 (pillar)	***62.5***	12.4	-	-	-	***94.3***	***52.6***	3.7	16.4
	支柱 (prop, support)	6.2	***62.2***	9.8	-	***100***	5.7	29.3	-	5.0
	円柱 (cylinder)	-	-	-	-	-	-	11.1	-	-
	列 (line, array)	17.5	22.1	-	3.0	-	-	2.6	***67.8***	***64.5***
	ライン (line)	3.8	2.7	15.8	***97.0***	-	-	4.0	4.4	-
	コラム (newspaper column)	4.5	-	***60.3***	-	-	-	-	9.4	-
	欄 (section, blank)	3.4	-	8.4	-	-	-	-	13.1	9.7
culture	培養 (growing of bacteria)	***70.9***	16.4	-	***100***	-	-	-	-	-
	栽培 (growing of plants)	22.4	***76.9***	-	-	-	-	-	-	-
	養殖 (raising of animals)	5.4	6.7	-	-	-	-	-	-	-
	訓練 (training)	-	-	-	-	-	-	***100***	***63.1***	-
	教育 (education)	-	-	-	-	-	-	-	36.9	
nail	釘 (fastener)	***79.3***	***79.2***	22.8	-	-	***96.6***	***92.4***	-	-
	爪 (body structure)	20.7	20.8	***77.2***	-	-	3.4	7.6	***100***	***100***
plant	植物 (flora)	***46.5***	***88.3***	***31.8***	***56.8***	-	***67.2***	-	-	-
	植木 (garden plant)	-	5.0	-	-	-	-	-	-	-
	プラント (industrial plant)	21.1	-	31.1	2.8	-	-	***85.7***	***81.6***	21.5
	装置 (instrument, device)	-	-	-	22.5	***87.9***	12.4	5.5	3.0	***46.2***
	工場 (factory, works)	-	-	8.0	5.3	-	-	-	-	-
	設備 (apparatus, facilities)	26.4	2.7	28.8	12.6	12.1	14.4	8.9	9.4	29.9
	建物 (building)	-	-	-	-	-	5.5	-	-	-

* English translations other than the target word are given in parentheses.

† Italicized *RAW* values indicate the most major translations; a hyphen (-) means that *RAW* is less than 2.5%.

ALL	2.435	C	1.724	F	1.860
A	2.237	D	1.455	G	2.232
B	2.149	E	1.898	H	2.005

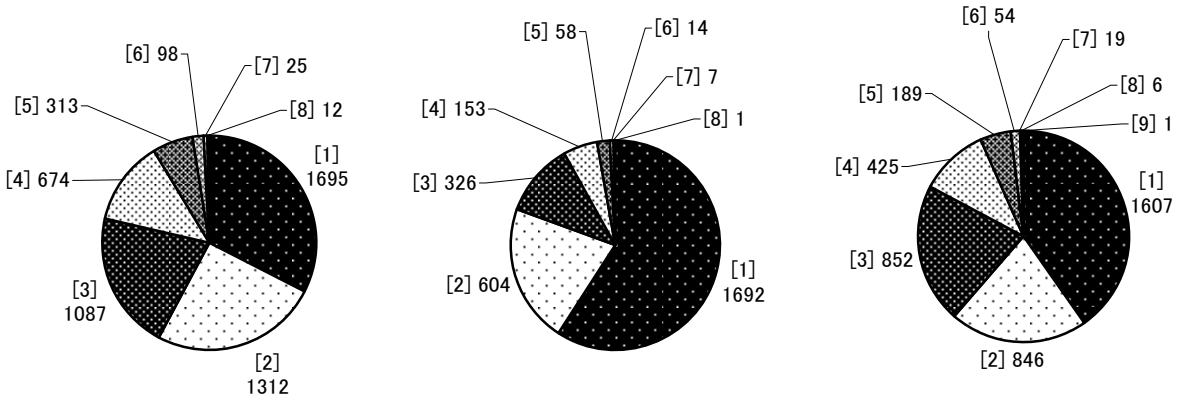
Figure 2 shows the distributions of the number of translations per target word in three cases: ALL, C, and G.

These data imply that division of the patent domain into subdomains generally makes lexical translation easier. In addition, the difficulty of lexical translation varies rather dramatically with subdomains; for example, lexical translation in the area of chemistry and metallurgy (C) seems much easier than that in the area of physics (G). Interestingly, the distribution of the number of translations per target word of the subcorpus G is similar to that of the whole corpus ALL. It indicates the necessity of dividing the area of physics into subareas. It should be noted that Section G in fact covers not only physics but also computer technologies.

3.4 Ratio of associated words suggesting the most major translation

The value of the ratio of associated words suggesting the most major translation, abbreviated to *RAW-MMT*, is very important from a practical point of view. That is, machine translation systems can fix a translation for a word with *RAW-MMT* exceeding a threshold. The optimum threshold depends on the performance of word-sense disambiguation or translation-word selection.

Figure 3 shows the distributions of *RAW-MMT* in



(a) ALL (b) C (Chemistry; metallurgy) (c) G (Physics)

Note: "[*m*] *n*" means that *n* target words have *m* translations.

Figure 2. Number of translations per target word.

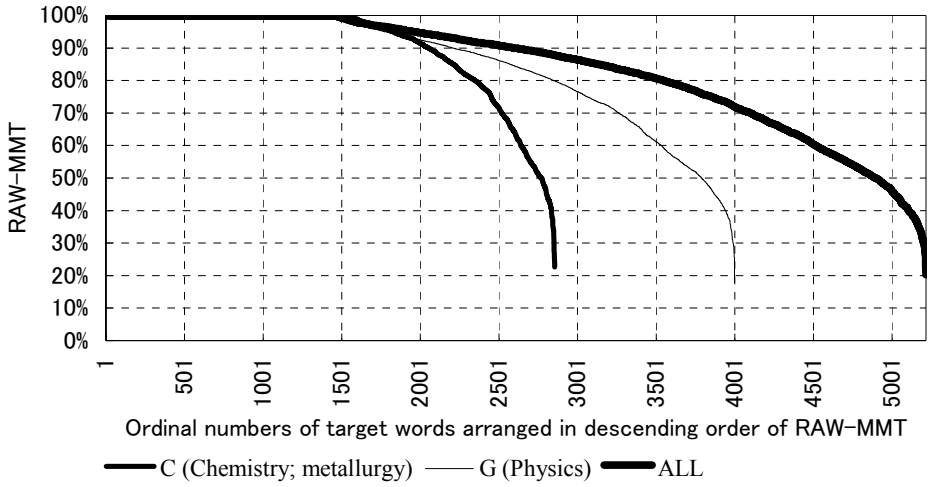


Figure 3. Ratio of associated words suggesting the most major translation.

three cases (ALL, C, and G). The horizontal coordinate represents the ordinal numbers of target words, which are arranged in descending order of *RAW-MMT*, and the vertical coordinate represents the value of *RAW-MMT*. For example, the curve for ALL shows that when the threshold for *RAW-MMT* is set at 90%, 80%, and 50%, translations can be fixed for 2,640 words, 3,591 words, and 4,946 words out of 5,261 words, respectively. The numbers of target words with *RAW-MMT* less than a threshold in the cases of both C and G are smaller than that for ALL. This indicates that division of the patent domain into subdomains makes lexical translation easier. Figure 3 also supports lexical translation in the area of chemistry and metallurgy being much easier than that in the area of physics.

4 Conclusion

A method using bilingual comparable corpora to correlate translations with associated words has been used to examine the domain dependence of translations of nouns in patent translation. The usefulness of two metrics, i.e., the number of translations per word and the ratio of associated words suggesting the most major translation, has been demonstrated. The experimental results indicate the necessity and effectiveness of dividing the patent domain into subdomains and adapting a bilingual dictionary to subdomains. One remaining problem is how to determine the optimum set of subdomains. Another important research issue is to evaluate how well the ratio of associated words suggesting a translation approximates a translation probability.

5 Acknowledgments

This work was supported by the New Energy and Industrial Technology Development Organization (NEDO). We were permitted to use the patent-abstract corpus in the experiments by courtesy of Japan Patent Information Organization (JAPIO). We are grateful to the members of AAMT-JAPIO Special Interest Group on Patent Translation for their valuable comments.

References

Church, Kenneth W. and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1): 22-29.

EDR. 1990. Bilingual Dictionary, Technical Report TR-029, Japan Electronic Dictionary Research Institute.

Fung, Pascale and Kathleen McKeown. 1997. Finding terminology translations from non-parallel corpora. In *Proceedings of the 5th Annual Workshop on Very Large Corpora*, pages 192-202.

Fung, Pascale and Lo Yuen Yee. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics* / *the 17th International Conference on Computational Linguistics*, pages 414-420.

Kaji, Hiroyuki. 2004. Bilingual-dictionary adaptation to domains. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 729-735.

Kaji, Hiroyuki. 2005. Adapting a bilingual dictionary to domains. *IEICE Transactions on Information and Systems*, E88-D(2): 302-312.

Kaji, Hiroyuki and Toshiko Aizono. 1996. Extracting word correspondences from bilingual corpora based on word co-occurrence information. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 23-28.

Kaji, Hiroyuki and Yasutsugu Morimoto. 2005. Unsupervised word-sense disambiguation using bilingual comparable corpora. *IEICE Transactions on Information and Systems*, E88-D(2): 289-301.

Rapp, Reinhard. 1995. Identifying word translations in non-parallel texts. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 320-322.

Rapp, Reinhard. 1999. Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 519-526.

Tanaka, Kumiko and Hideya Iwasaki. 1996. Extraction of lexical translations from non-aligned corpora, In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 580-585.