A Method of Augmenting Bilingual Terminology by Taking Advantage of the Conceptual Systematicity of Terminologies

Miki Iwai
Graduate School of
Interdisciplinary
Information Studies,
The University of Tokyo
mikii
@g.ecc.u-tokyo.ac.jp

Koichi Takeuchi Kazuya Ishibashi
Graduate School of
Natural Science and
Technology,
Okayama University
{koichi, ploi5t2g}
@cl.cs.okayama-u.ac.jp

Kyo Kageura
Graduate School of
Education,
The University of Tokyo
kyo@p.u-tokyo.ac.jp

Abstract

In this paper, we propose a method of augmenting existing bilingual terminologies. Our method belongs to a "generate and validate" framework rather than extraction from corpora. Although many studies have proposed methods to find term translations or to augment terminology within a "generate and validate" framework, few has taken full advantage of the systematic nature of terminologies. A terminology of a domain represents the conceptual system of the domain fairly systematically, and we contend that making use of the systematicity fully will greatly contribute to the effective augmentation of terminologies. This paper proposes and evaluates a novel method to generate bilingual term candidates by using existing terminologies and delving into their systematicity. Experiments have shown that our method can generate much better term candidate pairs than the existing method and give improved performance for terminology augmentation.

1 Introduction

In this paper, we propose a new way of generating new bilingual multi-word term pairs for augmenting existing bilingual terminologies.

There is growing demand for properly managed terminologies in many areas of society, e.g. in document authoring and management, in technical translation, in knowledge transfer and education, and in IR/NLP (Sager, 1990; Wright and Wright, 1997; Budin, 2008; Kockaert and Steurs, 2015). With the constant introduction of new terms in many domains, timely augmentation and update of terminologies is critical for proper terminology management, and automatic assistance for this process is greatly needed (Kockaert and Steurs, 2015). Many researchers have proposed various methods to augment terminologies automatically. As we will see in Section 2, these can be divided into two broad approaches, i.e. "extraction from corpora" approach and "generate and validate" approach. We focus on the latter approach, which fits better for augmenting or expanding *existing* terminologies, the task which is in strong demand in language industries but has not been much addressed from the NLP point of view.

A term in a terminology of a domain represents a concept of that domain. Majority of terms are complex in most domains in most languages. These complex terms represent concepts analytically, with each constituent element representing an important feature of the concept. A terminology, i.e. the set of terms of a domain, represents the structure of concepts of that domain more or less systematically. Although the extent of systematicity differ from language to language and from domain to domain, new terms are generally formed systematically within the conceptual system of the domain. If we can take into account this aspect of term formation for generating term candidates in the task of augmenting terminologies, we would be able to develop an effective way of help augmenting existing terminologies.

Against this backdrop, this paper proposes a new method of generating bilingual term candidates by taking advantage of the structural feature of terminology. The basic idea is as follows: define terminological network that reflects conceptual systematicity; identify "motivated" subnetworks within which term formation is supposed to be activated, and generate term candidates for each subnetwork.

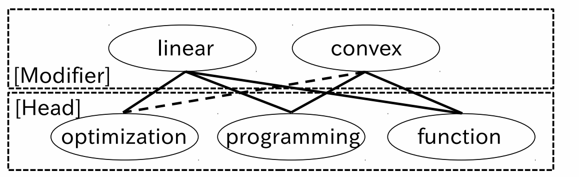

Figure 1: Example of generating a term candidate

The rest of this paper is organised as follows. Section 2 looks at related work and places the present work in context. Section 3 explains our proposed method. Sections 4 and 5 introduce the experimental setup and the results, respectively. Section 6 summarises the results and discusses remaining issues.

2 Related work

2.1 Automatic extraction/augmentation of bilingual terms/terminology

Bilingual term extraction from parallel or comparable corpora has been actively pursued since the 1990s (Dagan and Church, 1997; Fung and Mckeown, 1997; Gaussier, 1998; Chiao and Zweigenbaum, 2002; Kwong et al., 2004; Bernhard, 2006; Robitaille et al., 2006; Daille and Morin, 2008; Lefever et al., 2009; Laroche and Langlais, 2010), most of which use contextual information such as co-occurrence within aligned segments or contextual similarity. Research into the improvement of quality of corpora is also pursued (Morin et al., 2010; Li and Gaussier, 2010). The European project TTC (Terminology extracting Translation Tools and Comparable Corpora) is the culmination of this trend of research (Blancafort et al., 2010).

Some use the correspondence at the level of constituent elements of terms in finding term translations (Grefenstette, 1999; Tonoike et al., 2005; Tonoike et al., 2006; Daille and Morin, 2008), i.e. they generate term candidates in target language by translating constituent elements and validate their existence. These studies partly adopt the "generate and validate" framework. Sato et al. (2013) generated multi-word term pairs as bilingual term candidates by considering all possible pairs of constituent elements of terms in a terminology. The generated pairs are then validated by using web documents.

Our method adopts this "generate and validate" framework. More specifically, we take Sato et al. (2013) as a point of departure as the aim of this work is the same as the present work, i.e. extending existing bilingual terminologies. The method proposed by Sato et al. (2013) takes advantage of a general tendency that if one term is a compound, a part of the term is a term and a part of the term can be changed. For example, if a terminological lexicon contains, "linear programming", "linear optimization", "linear function", "convex programming" and "convex function", they can expect that the term "convex optimization" exists, even if this term is not listed in the lexicon. They generate term candidates consisting of two constituents by defining head-modifier bipartite graph and interpolate missing edges. Figure 1 shows this idea graphically.

The problems we identify with their method are (a) if applied straightforwardly, a huge number of bilingual term candidates are generated, and (b) the Kernighan-Lin algorithm they adopted (Kernighan and Lin, 1970) to partition head-modifier bipartite graph in order to reduce term candidates does not reflect systematic structure of terminologies. Following theoretical research in terminology (Sager, 1990; Kageura, 2002), we understand that new terms are formed within the conceptual-terminological subsystem surrounding the new concepts. So our main task is concerned with consolidating these subsystems consisting of tightly-related or "motivated" terms/concepts within which new terms are formed.

2.2 Structural nature of terminology

Terminologies in most languages contain a substantial number of complex terms (Cerbah, 2000; Nomura and Ishii, 1989). Research has shown that complex terms tend to show conceptual relationships systematically, with each constituent element representing an important feature of concepts represented by terms (Felber, 1984; Sager, 1990; Kageura, 2002).

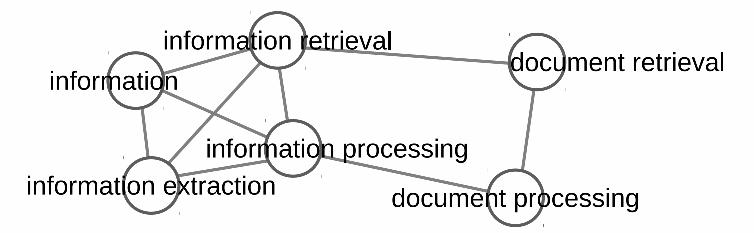

Figure 2: Terminology network of a putative terminology

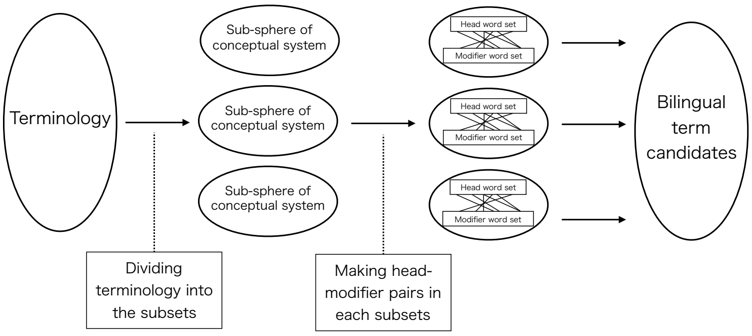
Terminology
Sub-sphere of conceptual system
Sub-sphere of conceptual system
Sub-sphere of conceptual system
Head word set
Modifier word set
Head word set
Modifier word set
Head word set
Modifier word set
Bilingual term candidates
Dividing terminology into the subsets
Making head-modifier pairs in each subsets

Figure 3: Proposed method

Kageura (2012) examined the systematic nature of terminologies by introducing terminological network, the vertices of which are terms and the edges of which consist of common constituent elements between terms. For instance, a putative terminology consisting of six terms, “information”, “information retrieval”, “information extraction”, “document retrieval”, “document processing”, and “information processing” makes a network as shown in Figure 2.

Kageura and Abekawa (2007) applied partitive clustering over the terminological network to obtain sub-groups of terminologies. Asaishi and Kageura (2011) comparatively analysed the formal nature of terminological structure by defining terminological networks of English and Japanese bilingual terminologies of several domains. Iwai et al. (2016) have shown that there is a reasonable amount of cross-lingual correspondence between sub-groups of English and Japanese terms identified by using community detection algorithms over the terminological network. As stated above, to identify conceptual subsystems consisting of closely-related concepts in terminology constitutes an essential part of our method. Iwai et al. (2016) showed that meaningful conceptual subsystems can be identified and extracted by applying relevant network partition algorithms to terminological networks.

3 Method of term candidate generation

Starting from a given terminology, the method of term candidate generation we propose consists of two steps:

1. Dividing a terminology into subgroups each of which consists of terms representing closely related concepts; and

2. Generating bilingual term candidates by generating possible combinations of constituent elements of terms included in each subgroup.

Figure 3 shows an outline of the method.

While Sato et al. (2013) firstly considered all possible head-modifier pairs for all terms in terminology and then reduced the number of term candidates by applying Kernighan-Lin algorithm to the head-modifier bipartite graph, our method first consolidate subgroups of terms and generate term candidates for each subgroup separately. Note that this is not just a methodological alternative, but reflects theoretical understanding of how new terms are formed, as stated above.

3.1 Identifying "motivated" sub-groups of terms

We first construct terminological networks (Kageura, 2012), and then apply partitive clustering or community detection algorithm to the network. This manipulation identifies motivated sub-groups of terms within a given terminology. As terms are formed within subsystems of concepts, this serves for reducing the number of generated term candidates while at the same time increasing the plausibility of candidates. After dividing terminological networks into sub-groups or clusters, we generated a head-modifier set for each cluster.

The steps for this process are as follows:

1. Decompose each term into its constituent elements;
2. Generate terminological network with terms as vertices and common constituent elements as edges;
3. Divide the generated terminological network into clusters using a community detection algorithm.

For step 1, we used MeCab[1] with UniDic[2] to decompose Japanese terms into constituent elements. For English terms, we decompose terms using spaces and other punctuations and then apply stemming and lemmatisation of constituent words using a lemmatiser[3]. Although POS taggers, such as Stanford POS Tagger[4], are widely used for pre-processing English sentences or phrases, we used here the lemmatiser because (a) our aim is to extract semantically identical units by removing inflectional (and sometimes derivational) variations and (b) we do not need POS-information. Previous work has shown that approximately matching units can be extracted for English and Japanese terminologies by applying these pre-processing steps (Asaishi and Kageura, 2011).

For step 2, we used python igraph library[5] to generate terminology networks for English and Japanese. We removed functional words (symbols, numbers, prepositions and articles for English; symbols, numbers, particles and auxiliary verbs for Japanese) as they do not represent conceptual characteristics.

For step 3, we adopted Potts spin glass algorithm to divide the terminology networks into clusters. Many community detection algorithms have been proposed (Clauset et al., 2004; Rosvall and Bergstrom, 2008; Raghavan et al., 2007; Blondel et al., 2008; Pons and Latapy, 2006; Newman, 2006). After examining several commonly used methods, we decided to adopt Potts spinglass-based method (Reichardt and Bornholdt, 2006), which works by solving the global optimization problem (Kirkpatrick, 1984). Not only is this method reported to work well in several experiments, the underlying concept reflects nicely the task of extracting motivated sub-groups of terminologies (Kageura and Abekawa, 2007).

3.2 Generating bilingual term candidates

After obtaining clusters on sub-groups of terms, we generated bilingual term candidates as follows:

1. Identify corresponding English and Japanese terms contained in each cluster. As English and Japanese clusters do not match completely (Iwai et al., 2016), we generated term candidate pairs in three different ways in step 1: (a) based on Japanese clusters (Japanese), (b) based on English clusters (English), and (c) based on the intersections of Japanese and English clusters (mix).
2. Generate bilingual pairs of constituent elements (henceforth constituent pairs). This is carried out first by identifying single-word term pairs and then subtracting them from multi-word terms and making remaining elements as pairs recursively.
3. Generate head-modifier pairs for constituent elements of source language terms, as shown in Figure 4. We identify head-modifier relations by identifying constituents on the left as modifiers and on the right as heads, as English (and Japanese, for that matter) complex terms are head final. We also assumed that if a term constitutes more than three words, two constituent elements can replace as one semantics unit. For example, we can consider that "data" is the modifier and "processing

[1] http://code.google.com/p/mecab/

[2] http://pj.ninjal.ac.jp/corpus_center/unidic

[3] http://www.nltk.org/api/nltk.stem.html

[4] http://nlp.stanford.edu/software/tagger.shtml

[5] http://igraph.org

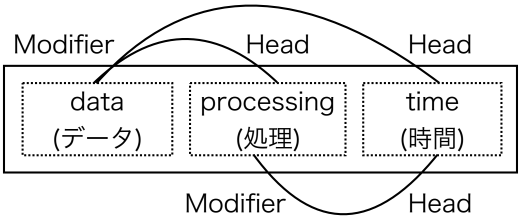

Figure 4: Extracting head-modifier pairs

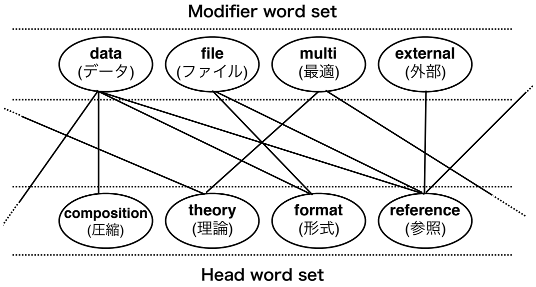

Figure 5: Example of bipartite graph

Dom.	Lang.	T	1	2	3	4+
Com.	En	16259	2634 (16.20%)	9044 (55.62%)	3645 (22.42%)	936 (5.76%)
	Ja	16259	2002 (12.31%)	7141 (43.92%)	4782 (29.41%)	2334 (14.36%)
Ecn.	En	9120	1219 (13.37%)	4858 (53.27%)	1659 (18.19%)	1384 (15.17%)
	Ja	9120	947 (10.38%)	3753 (41.15%)	2814 (30.86%)	1606 (17.61%)

Table 1: The distribution of terms in each terminology

time" is the head in Figure 4. However, we considered only head-modifier pairs by minimum unit in this time. We set English as source language for convenience of processing; there is no inherent technical reason for us to make the process directional in terms of languages.

4. Generate a bipartite graph based on the head-modifier pairs of the source language, as shown in Figure 5.
5. Take the direct product of the head and modifier vertices to generate extended head-modifier pairs from that bipartite graph.
6. Create new bilingual term pairs by taking translations for each constituent elements of the head-modifier pairs using constituent pairs.

The candidate term pairs generated through this process are then validated using web documents.

4 Experimental setup

4.1 Seed terminologies

For evaluation, we used two terminological dictionaries, i.e. one in the field of computer science (Aiso, 1993) and the other in the field of economics (Yuhikaku, 1986). These are two of the five terminological dictionaries used in Sato et al. (2013). Table 1 shows the number and ratio of terms by length in each terminology, i.e. single terms, terms with two constituents, terms with three constituents and terms with four or more constituents. "Dom." stands for domain, "Lang." stands for language, and "T" indicates the number of terms. From Table 1, we can observe that these terminologies contain many complex terms.

4.2 Terminological network and candidate generation

We constructed terminological networks for English and for Japanese separately for these two datasets. Table 2 shows the quantitative nature of the terminological networks, in which N stands for the number of constituent elements, V the number of vertices, E the numbers of edges, and S the number of isolated terms. We can observe that each network consists of a single giant component (max subgraph) and several small components (others) including isolated vertices.

We then extracted max subgraph and divided it into clusters. The number of clusters was set in two ways, i.e. 25 and 10. These numbers were decided heuristically, referring to the number of subdomains listed in handbooks and in academic societies. The number of candidates generated from these clusters is given in Table 3, which also provides the number of candidates generated from the method by Sato et al. (2013). Note that our method produces smaller number of term candidates.

Dom.	**Lang.**	T	V	E	S	**max subgraph**	
						V	E
Com.	En	16259	14186	992319	1100	13046	992293
	Ja	16259	15062	998245	1468	13380	997034
Ecn.	En	9120	8922	278836	749	8127	278812
	Ja	9120	4647	267603	863	8096	26784

Table 2: Basic quantities of terminologies and terminological networks

Dom.	**10clusters**			**25clusters**			**Sato et al. (2013)**
	En	Ja	mix	En	Ja	mix	
Com.	106,422	37,741	27,252	93,175	27,342	20,891	202,446
Ecn.	33,348	12,478	10,009	29,075	9,112	7,885	82,806

Table 3: The number of generated term candidates

4.3 Collecting web documents for validation

Web documents are collected separately for two languages and stored in a database. To avoid collecting irrelevant web pages, we used domain keywords (the name of the domain such as “computer science”) together with individual terms for collecting documents.

Web documents for computer science were collected in October and November 2014, by using terms and the domain keywords “computer science” (English) and “情報科学” (“information science” for Japanese) (see 3.1). Web documents for economics were collected at the end of December 2014, with domain keywords “economics” (English) and “経済学” (“economics” for Japanese). Table 4 shows the basic quantities of the collected documents. We extracted 200 pages randomly from the English data and manually checked the number of technical documents. The result is shown in Table 5. Approximately 60 % of the documents were technical in both domains.

5 Evaluation

We evaluated our method in two ways. First, we compared our result with Sato et al. (2013) in terms of the number of retained candidates after validation. Second, to evaluate precision, we extracted top 100 candidates ranked according to (a) the sum of English and Japanese occurrences and (b) the Jaccard coefficient. Note that we do not make comparison between our approach and the approach of extracting terms from corpora, because their experimental setups are very different to each other.

5.1 Comparison of the number of retained candidates after validation

The candidate term pairs generated in six different ways (two cluster sizes of 10 and 25 by based on Japanese clusters, based on English clusters, and based on the intersections of Japanese and English clusters) were validated by 2 steps using the web documents (see 4.3).

1. Searching bilingual term candidates from collected web documents and retaining candidate pairs of which both English part and Japanese part occur at least once in the documents.

2. Calculating a Jaccard coefficient by using retained candidate pairs.

In step 1, instead of using the web search directly, we first pool the web documents relevant to the two domain. It is to avoid repeatedly searching the web for every candidate pairs. In step 1, we validate English and Japanese terms separately, as we can assume that the candidates are aligned. However, it is still useful to validate the bilingual co-occurrences in the web documents. In order to observe that, we used Jaccard coefficient.

Dom.	English	Japanese	total
Com.	121,740	43,868	165,608
Ecn.	98,630	58,411	157,040

Table 4: The number of collected web documents

Dom.	Technical documents	percentage
Com.	126	63.0%
Ecn.	130	65.0%

Table 5: Percentage of technical documents

Dom.	10clusters			25clusters			Sato et al. (2013)
	En	Ja	mix	En	Ja	mix	
Com.	39,198 (36.83%)	17,239 (45.68%)	13,583 (49.84%)	34,683 (37.22%)	14,123 (51.65%)	11,628 (55.67%)	9,849 (4.87%)
Ecn.	12,105 (36.30%)	6,718 (52.70%)	5,957 (59.52%)	10,862 (37.36%)	5,707 (62.63%)	5,227 (66.29%)	6,523 (7.88%)

Table 6: The result of validation (filtering)

5.1.1 Filtering by using collected web documents

We first did the filtering by using collected web documents to reduce the number of generated bilingual term candidates. Candidate pairs of which both English part and Japanese part occur at least once in the corpus were retained as validated terms. Table 6 shows the result. The first line in each domain shows the number of validated candidates. The second line shows their percentage against the number of candidate pairs given in Table 3. It shows that the number of terms retained after validation is generally larger in our methods than Sato et al. (2013), with exceptions ("mix" for 10 clusters, and "Ja" and "mix" for 25 clusters in economics). In all cases, the ratio of retained candidates is much higher in our method than Sato et al. (2013). These results indicate that our proposed method:

- performs both more effectively in terms of computational cost and in terms of recall, assuming that the validated terms have roughly the same level of pairing precision and termhood precision; and
- enables us to control the balance between recall and precision, by changing the number of clusters as well as the pairing methods.

The first point indicates that our method successfully captures the conceptual subsystems/terminological subgroups within the dynamics of which new terms are formed. The second point shows that our method gives us applicational flexibility.

5.1.2 Calculating Jaccard coefficient

After filtering by collected web documents, we searched retained bilingual term candidates with search engine and calculated Jaccard coefficient by using the number of hit. In order to keep the comparison with Sato et al. (2013) sensible, we chose the validated candidates generated from "mix" for 25 clusters, as the number of validated terms in the two domains is close to that by Sato et al. (2013) (although "Ja" pairing for 10 clusters is the closest in economics, we chose the same setting for the two domains). Jacard coefficient is defined as:

$$Jaccard(L1, L2) = \frac{H(L1) \wedge H(L2)}{H(L1) \vee H(L2)} = \frac{H(L1) \wedge H(L2)}{H(L1) + H(L2) - H(L1 \wedge L2)},$$

where $L1$ and $L2$ indicate English and Japanese parts (or vice versa) of a candidate pair in our case, and $H(x)$ is the number of documents in which they occur. If the number of hits is zero, the Jaccard coefficient is defined to be zero. In filtering by using collected web documents, the process retained candidate pairs that either English part or Japanese part occur. Therefore, it is considered that non-parallel candidate pairs are retained. By calculating Jaccard coefficient with the number of hit in search engine and retaining candidate pairs that Jaccard coefficient is positive, we finally extract candidate pairs that is validated parallel. We used Bing search API as search engine. Table 7 shows the result of the total

Dom.	"mix" for 25 clusters	Sato et al. (2013)
Com.	9,471 (81.45%)	2,261 (23.00%)
Ecn.	4,707 (90.05%)	2,286 (35.05%)

Table 7: The result of total number of positive Jaccard coefficients

	Occurrences			Jaccard		
Dom.	pairing	term	partial	pairing	term	partial
Com.	82 (61)	56 (28)	16 (17)	86 (89)	72 (51)	13 (15)
Ecn.	87 (56)	69 (37)	24 (16)	95 (91)	86 (60)	8 (18)

Table 8: Precision of top 100 candidates

number of candidate term pairs that take positive values of Jaccard coefficients. The result indicates that our method generates many more potentially valid candidate pairs than the method by Sato et al. (2013).

5.2 Precision of top 100 candidates

The top 100 candidates generated by "mix" for 25 clusters, ranked according to the sum of English and Japanese occurrences and to the Jaccard coefficient, were manually evaluated for each domain. The evaluation was carried out from two points of view, i.e. (a)whether the Japanese and English matches or not (pairing), and (b)whether the Japanese candidates can be regarded as a term in the domain in question (term). For (b), we also counted partial-terms (partial). The evaluation was carried out by one of the authors. Table 8 shows the result, together with the corresponding results given in Sato et al. (2013) (in bracket). Table 8 shows that except for "pairing" by Jaccard in computer science, our method is consistently better than Sato et al. (2013) in terms of precision as well.

6 Conclusion and future work

In this paper we proposed a method of augmenting existing bilingual terminological lexicon. We introduced a way of generating candidate term pairs which reflect the conceptual system/terminological group within which new terms are formed, by taking advantage of the "motivated" structure of terminologies. Compared with the method proposed so far, our method consistently shows higher performance, which indicates that our method succeeded in identifying, to a reasonable extent, the conceptual subsystem/terminological subgroups within which terms are formed. The method also has more applicational flexibility.

We are currently addressing the following issues:

- Extending our method so that it can generate and validate terms with more than three constituent elements. For example, if a term consists of more than three words, it is natural to decompose it into 2 words as one unit and the other one word from the point of semantic structure. In this way, we try to apply generating bilingual term candidates that consists of more than three words.
- Improving the pairing module. As of now, we examined English as source language and Japanese as target language. However, we can consider reverse pattern in our proposed method. Directional property of language and correspondence of translation words are one of the points of that we need to address in the future.
- Analysing non-validated candidates (error analysis). Now that it was shown that the proposed method can capture, to a reasonable extent, conceptual subsystem within which new terms are generated, it is important to analyse non-validated candidates to obtain further insights into candidate generation process.
- Finding a way of suggesting reasonable number of clusters. As can be inferred from Tables 4 and 7, the best number of clusters may differ from domain to domain.

In addition, we are planning to extend our research into the following directions:

- Applying our method to different language pairs. We are planning to apply our method to Chinese-English and Korean-English pairs.
- Clarifying the difference between the "generate and validate" framework and extraction from parallel or comparable corpora. Although the comparison of these two approaches are difficult, because not only the theoretical assumption and the range of relevant applications but also the range of data which can be used differ greatly (the "generate and validate" approach in general can use wider variety of data as they are used for validation rather than sources from which terms are extracted), it would still be interesting to examine the relationship between these two approaches on the empirical basis.

Acknowledgements

This work is partly supported by JSPS Grant-in-Aid (A) 25240051 "Archiving and using translation knowledge to construct collaborative translation training aid system."

References

Hideo Aiso. 1993. *Dictionary of information processing*. Tokyo: Ohm.

Takuma Asaishi and Kyo Kageura. 2011. Comparative analysis of the motivatedness structure of Japanese and English terminologies. In *Proceedings of the 9th International Conference on Terminology and Artificial Intelligence (TAI)*, pages 38–44.

Delphine Bernhard. 2006. Multilingual term extraction from domain-specific corpora using morphological structure. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 171–174.

Helena Blancafort, Béatrice Daille, Tatiana Gornostay, Ulrich Heid, Claude Méchoulam, and Serge Sharoff. 2010. TTC: Terminology extraction, translation tools and comparable corpora. In *Proceedings of the 14th European Association for Lexicography (EURALEX) International Congress*, pages 263–268.

Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008.

Gerhard Budin. 2008. Global content management. *Topics in Language Resources for Translation and Localisation*, pages 121–134.

Farid Cerbah. 2000. Exogeneous and endogeneous approaches to semantic categorization of unknown technical terms. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING)*, pages 145–151.

Yun-Chuang Chiao and Pierre Zweigenbaum. 2002. Looking for candidate translational equivalents in specialized, comparable corpora. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING)*, volume 2, pages 1–5.

Aaron Clauset, Mark E. J. Newman, and Cristopher Moore. 2004. Finding community structure in very large networks. *Physical Review*, 70:66–111.

Ido Dagan and Ken Church. 1997. Termight: Cordinating humans and machines in bilingual terminology acquisition. *Machine Translation*, 12:89–107.

Béatrice Daille and Emmanuel Morin. 2008. Effective compositional model for lexical alignment. *In Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP)*, pages 95–102.

Helmut Felber. 1984. *Terminology manual*. UNESCO, Paris.

Pascale Fung and Kathleen Mckeown. 1997. Finding terminology translations from non-parallel corpora. In *Proceedings of 5th International Workshop of Very Large Corpora (WVLC-5)*, pages 192–202.

Éric Gaussier. 1998. Flow network models for word alignment and terminology. In *Proceedings of the 17th International Conference on Computational Linguistics (COLING)*, pages 444–450.

Gregory Grefenstette. 1999. The world wide web as a resource for example-based machine translation tasks. In *Proceedings of the ASLIB Conference on Translating and the Computer*, volume 21.

Miki Iwai, Koichi Takeuchi, and Kyo Kageura. 2016. Cross-lingual structural correspondence between terminoogies: The case of English and Japanese. In *Proceedings of the 12th International conference on Terminology and Knowledge Engineering (TKE)*, pages 14–23.

Kyo Kageura and Takeshi Abekawa. 2007. Modelling and exploring the network structure of terminology using the Potts spin glass model. *In Proceedings of the 10th Conference of the Pacific Association for the Computational Linguistics (PACLING)*, pages 236–245.

Kyo Kageura. 2002. *The dynamics of terminology: A descriptive theory of term formation and terminological growth*. John Benjamins, Amsterdam.

Kyo Kageura. 2012. *The quantitative analysis of the structure and dynamics of terminologies*. Amsterdam: John Benjamins.

Brian W. Kernighan and Shunjiang Lin. 1970. An efficient heuristic procedure for partitioning graphs. *Bell Systems Technical Journal*, 49(2):291–307.

Scott Kirkpatrick. 1984. Optimization by simulated annealing: Quantitative studies. *Journal of Statistical Physics*, pages 975–986.

Hendrik J. Kockaert and Frieda Steurs, editors. 2015. *Handbook of terminology*, volume 1. John Benjamins, Amsterdam.

Oi Yee Kwong, Benjamin K. Tsou, and Tom B. Y. Lai. 2004. Alignment and extraction of bilingual legal terminology from context profiles. *Terminology*, 10(1):81–99.

Audrey Laroche and Philippe Langlais. 2010. Revisiting context-based projection methods for term-translation spotting in comparable corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, pages 617–625.

Els Lefever, Lieve Macken, and Veronique Hoste. 2009. Language-independent bilingual terminology extraction from a multilingual parallel corpus. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 496–504.

Bo Li and Eric Gaussier. 2010. Improving corpus comparability for bilingual lexicon extraction from comparable corpora. *In Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, pages 644–652.

Emmanuel Morin, Béatrice Daille, Koichi Takeuchi, and Kyo Kageura. 2010. Brains, not brawn: The use of "smart" comparable corpora in bilingual terminology mining. *ACM Transactions on Speech and Language Processing (TSLP)*, 7(1).

Mark E. J. Newman. 2006. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74(3):36–104.

Masaaki Nomura and Masahiko Ishii. 1989. *List of stems in Japanese technical terms*. Technical report, National Institute for Japanese Language and Linguistics, Tokyo.

Pascal Pons and Matthieu Latapy. 2006. Computing communities in large networks using random walks. *Journal of Graph Algorithms and Applications*, 10(2):191–218.

Usha Nandini Raghavan, Réka Albert, and Soundar Kumara. 2007. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, 76(3):36–106.

Joerg Reichardt and Stefan Bornholdt. 2006. Statistical mechanics of community detection. *Physical Review E*, 74(1):16–110.

Xavier Robitaille, Yasuhiro Sasaki, Masatsugu Tonoike, Satoshi Sato, and Takehito Utsuro. 2006. Compiling French-Japanese terminologies from the web. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 225–232.

Martin Rosvall and Carl T. Bergstrom. 2008. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 105(4):1118–1123.

Juan C. Sager. 1990. *A practical course in terminology processing*. Amsterdam: John Benjamins.

Koichi Sato, Koichi Takeuchi, and Kyo Kageura. 2013. Terminology-driven augmentation of bilingual terminologies. *In Proceedings of the XIV Machine Translation Summit (MT Summit)*, pages 3–10, 9.

Masatsugu Tonoike, Mitsuhiro Kida, Toshihiro Takagi, Yasuhiro Sasaki, Takehito Utsuro, and Satoshi Sato. 2005. Effect of domain-specific corpus in compositional translation estimation for technical terms. *Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP)*, pages 116–121.

Masatsugu Tonoike, Mitsuhiro Kida, Toshihiro Takagi, Yasuhiro Sasaki, Takehito Utsuro, and Satoshi Sato. 2006. A comparative study on compositional translation estimation using a domain/topic-specific corpus collected from the web. In *Proceedings of the 2nd Web as Corpus Workshop*, pages 11–18.

Sue Ellen Wright and Leland D. Wright Jr.. 1997. Terminology management for technical translation. *Handbook of Terminology Management*, 1:147–159.

Yuhikaku. 1986. *Dictionary of economy terms*. Yuhikaku, Tokyo.