Terminology-driven Augmentation of Bilingual Terminologies

Koichi Sato **Koichi Takeuchi**
Graduate School of Natural Science and Technology
Okayama University
koichi@cl.cs.okayama-u.ac.jp

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

Abstract

This paper proposes a way of augmenting bilingual terminologies by using a "generate and validate" method. Using existing bilingual terminologies, the method generates "potential" bilingual multi-word term pairs and validates their status by searching web documents to check whether such terms actually exist in each language. Unlike most existing bilingual term extraction methods, which use parallel or comparable corpora, the proposed method can take advantage of a wider variety of textual corpora. Experiments using Japanese-English terminologies of five domains show that the method is highly promising.

1 Introduction

In this paper we propose a way of detecting new bilingual term pairs for augmenting bilingual terminologies by using a "generate and validate" method. Augmenting bilingual terminologies is *sine qua non* for terminology managers, translators and document managers (Sager, 1990), and its importance is growing in accordance with the rapid growth of terminologies in many domains.

In general, new terms they tend to be created in a systematic way by compounding (Sager, 1990; Ananiadou, 1994; Justeson and Katz, 1995; Cerbah, 2000; Kageura, 2012), resulting in an abundance of multi-word terms (MWTs). This fact results in a tendency for the correspondences between constituent elements to be retained across languages to a substantial extent.

This provides us with a chance to take advantage of the information contained in existing terminologies to augment and enrich terminologies with new terms, based on a simple idea: If a terminological lexicon contains "linear programming," "linear optimization," "linear function," "convex programming" and "convex function," we can reasonably assume that the term "convex optimization," which is not listed in the terminology, may, or will come to, exist (Figure 1). By generating "potential" term candidates and validating their existence by using web data, it should be possible to identify a range of new terms which are not covered in existing terminologies. Assuming bilingual correspondence at the level of constituent units of terms, it is possible to extend this idea to obtain new bilingual term pairs. Based on this idea, we developed a fully operating system for detecting new bilingual term pairs in order to augment bilingual terminologies.

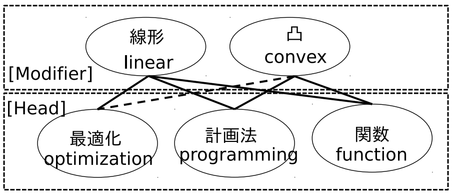

Figure 1: Existing and "potential" term pairs

The paper is organised as follows. Section 2 briefly looks at related work. Section 3 explains the system arrangement and the methods and algorithms adopted in the modules of the system. In particular, we detail the graph-based generation of term candidate pairs. Experimental results are introduced and discussed in section 4. Section 5 discusses remaining issues. Except for section 2, Japanese-English language pairs are assumed,

2 Related work

Since the 1990s, bilingual term extraction from parallel or comparable corpora has been actively pursued (Dagan and Church, 1997; Fung and McKeown, 1997; Gaussier, 1998; Chiao and Zweigenbaum, 2002; Kwong et al., 2004; Tonoike et al., 2005; Bernhard, 2006; Robitaille et al., 2006; Daille and Morin, 2008; Lefever et al., 2009;

Sima'an, K., Forcada, M.L., Grasmick, D., Depraetere, H., Way, A. (eds.)
Proceedings of the XIV Machine Translation Summit (Nice, September 2–6, 2013), p. 3–10.

Laroche and Langlais, 2010; Li and Gaussier, 2010; Morin et al., 2010).

Although extracting bilingual term pairs from parallel corpora generally attains higher precision than extracting them from comparable corpora, the problem of limited availability of parallel corpora has led to a great deal of research into bilingual term extraction using comparable corpora. In addition to work that has resulted in the steady improvement of algorithms, there are studies that address improvement of corpus comparability (Morin et al., 2010; Li and Gaussier, 2010). In the EU, research into corpus-based term extraction culminated in an EU project (TTC, 2012).

While the essential information explored in these methods is the correspondence between two languages (most typically aligned segments in the case of parallel corpora and degree of correspondence between context vectors in the case of comparable corpora), some have taken advantage of the abundance of MWTs and used the translational relationships between constituent units of MWTs (Tonoike et al., 2005; Daille and Morin, 2008). They partially take a "generate and validate" approach, for detecting target language expressions, although the essential framework is still oriented to "extraction."

These corpus-based approaches have shown steady technical advancement and improvement, but the results are essentially restricted by available corpora and not anchored to existing terminologies. From the point of view of augmenting terminologies for terminological management, more "terminology-driven" methods, i.e. those that make use of existing terminologies, are required. Our method takes this approach; it is complementary to existing work.

3 System and methods

3.1 Overall framework of the system

The system consists of three main modules:

(a) the module that generates potential term candidate pairs;

(b) the module that collects a set of web documents against which the existence of term candidate pairs is validated;

(c) the module that validates and ranks term candidate pairs.

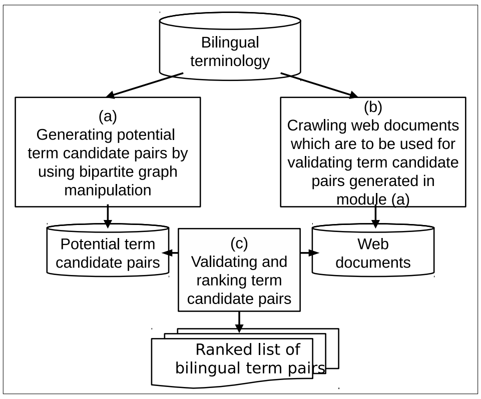
Bilingual terminology
(a) Generating potential term candidate pairs by using bipartite graph manipulation
(b) Crawling web documents which are to be used for validating term candidate pairs generated in module (a)
Potential term candidate pairs
(c) Validating and ranking term candidate pairs
Web documents
Ranked list of bilingual term pairs

Figure 2: Main modules of the system

Figure 2 shows the main modules of the system.

Among these, module (a) constitutes the core part of our approach. Validating (and ranking) candidate pairs generated in module (a) constitutes an essential part for our method. Currently we use the web as a source against which the existence of candidate term pairs is validated because it contains many new terms, but other sources could also be used for validation. The current system configuration is such that relevant documents are crawled from the web in advance, but it would also be possible to dynamically throw generated term candidate pairs into the web search engine for validation. We did not take this approach for reasons related to search engine api and in order to controlling evaluation and diagnosis.

3.2 Generating term candidate pairs

The following steps are carried out to generate term candidate pairs:

1. Decompose MWTs into components (CUs);
2. Establish correspondences between source language terms (SLT; Japanese in the present context) CUs and target language terms (TLT; English) CUs;
3. Generate head-modifier pairs for SLT CUs;
4. Generate a bipartite graph based on SLT head-modifier pairs;
5. Partition the bipartite graph;
6. For each connected component of the bipartite graph, take the direct product of the head and modifier vertices to generate extended head-modifier pairs;

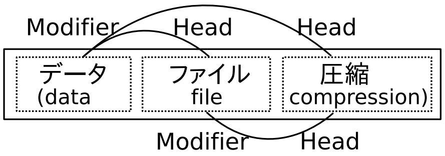

Figure 3: Extracting head-modifier pairs from an MWT "data file compression"

7. For each newly created SLT head-modifier pair, take the corresponding TLT CUs and generate corresponding TLT head-modifier pairs, then generate paired MWTs.

For step 1, MeCab[1] and Stanford POS Tagger (Toutanova et al., 2003)[2] are used for decomposing terms and POS-tagging CUs for Japanese and English, respectively. We retained content units, and functional units directly attached to them. For step 2, we start from aligned CU pairs taken from simple term pairs, and extend aligned pairs by iteratively removing aligned pairs from MWT pairs that have the same number of SLT CUs and TLT CUs.

In the third step, head-modifier pairs are generated for SLT CUs, using the fact that Japanese MWTs are head final. We extract all possible head-modifier pairs from each MWT, as shown in Figure 3. From an MWT with N constituent elements, $\binom{N}{2}$ head-modifier pairs are extracted.

Algorithm 1 Construction of bipartite graoph

```
Input: $Terms$: A set of CU sequences of terms
Input: $AlignedPairs$: A set of bilingually aligned CUs
Output: $TermGraph$: Bipartite graph with CUs as vertices
$TermCandidates \leftarrow \emptyset$
$Heads \leftarrow Modifiers \leftarrow Edges \leftarrow \emptyset$
$TermGraph \leftarrow (Heads, Modifiers, Edges)$
$SeedTerms \leftarrow Terms$ all CUs of which are in $AlignedPairs$
for all $SeedTerm \in SeedTerms$ do
    $N \leftarrow |SeedTerm|$
    for $i \leftarrow \{1, ..., N-1\}$ do
        $Modifier \leftarrow SeedTerm[i]$
        $Modifiers \leftarrow Modifiers \cup \{Modifier\}$
        for $j \leftarrow \{i+1, ..., N\}$ do
            $Head \leftarrow SeedTerm[j]$
            $Heads \leftarrow Heads \cup \{Head\}$
            $Edges \leftarrow Edges \cup \{(Head, Modifier)\}$
        end for
    end for
end for
return TermGraph
```

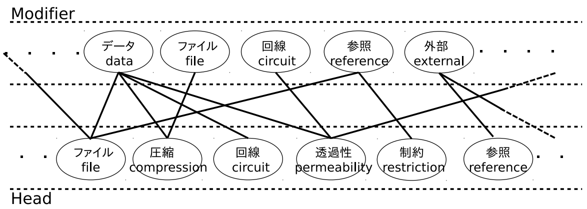

Figure 4: The head-modifier bipartite graph

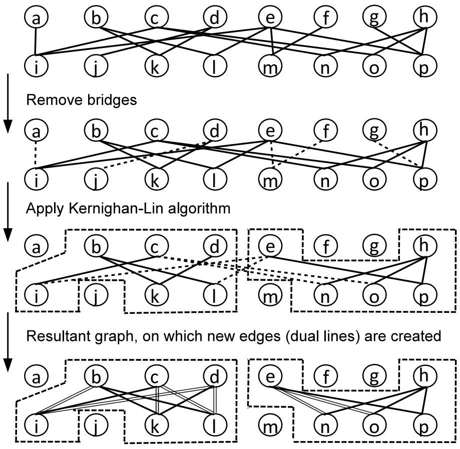
a
b
c
d
e
f
g
h
i
j
k
l
m
n
o
p
Remove bridges
Apply Kernighan-Lin algorithm
Resultant graph, on which new edges (dual lines) are created

Figure 5: Partitioning the bipartite graph

In step 4, the bipartite graph (as shown in Figure 4) is constructed from a set of head-modifier pairs obtained in step 3. Algorithm 1 shows the procedure in pseudo-code. The bipartite graph is generated by using only those SLT CUs which have corresponding TLT CUs (given in step 2).

Taking the direct product of the head and modifier vertices for this "raw" bipartite graph would generate a great number of head-modifier pairs which are not likely to be possible terms, due to the existence of unmotivated bridges. Assuming that there are reasonable coherent sub-graphs that contain potential term pairs, we thus partition the bipartite graph to create components of a reasonable size in step 5. This is done by: (a) first removing bridges from the graph, and (b) then partitioning large components by using the Kernighan-Lin algorithm (Kernighan and Lin, 1970). The Kernighan-Lin algorithm is a heuristic algorithm for partitioning connected components of a graph into two connected components of similar size.

Figure 5 illustrates the process of partitioning the graph according to this procedure. A wider range of methods could potentially be applied to this step.

[1] http://code.google.com/p/mecab/
[2] http://nlp.stanford.edu/software/tagger.shtml

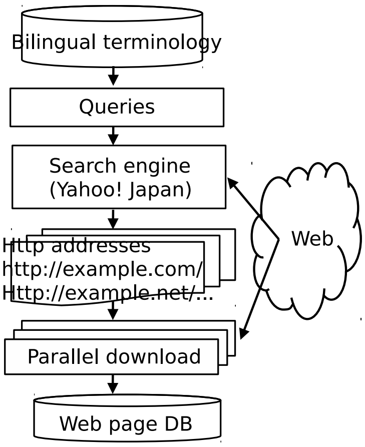

Figure 6: Collecting web documents

In step 6, the direct product of the head and modifier vertices is taken for each component, generating potential SLT candidates. Finally, in step 7 term candidate pairs are generated by taking and concatenating corresponding TLT CUs.

3.3 Collecting web documents

Both SL and TL web documents are collected, by using SLTs and TLTs listed in terminology as a query to the search engine (Figure 6). To avoid collecting irrelevant documents, therefore, we combined domain keywords (the name of the domain itself, such as "computer science," for example) with each query term. The top 20 documents are collected for each query. As a search engine, we currently use the Yahoo! Japan api[3]. Parallel downloading is carried out to improve speed. The obtained documents are stored using Groonga[4], which provides efficient full text search functions.

3.4 Validating term candidate pairs

The generated potential term candidate pairs are validated against the web documents, and the pairs for which both the SLT candidate and the TLT candidate occur at least once in the documents are retained. Currently, the result can be ranked in accordance with the number of occurrences of either SLT or TLT candidates, their average, or according to Jaccard similarity coefficient between SLT and TLT, which is defined as:

$$Jaccard(SLT, TLT) = \frac{H(SLT \wedge TLT)}{H(SLT \vee TLT)}$$

where H is the number of hits of the term in the document set.

Note that ranking by the number of SLT or TLT candidate hits provides information related to whether or not the candidate is likely to be a valid term, while the ranking by Jaccard similarity coefficient measures how likely it is that the SLT and TLT candidates are actually a corresponding pair. In the proposed framework, using the Jaccard coefficient can be regarded as redundant, as the correspondence between SLT and TLT candidates is kept by CU level correspondences. We still find it important to use the Jaccard coefficient as evaluating experimental results by using Jaccard coefficient enables us to see to what extent we can rely on separate and independent validation for SLT and TLT candidates, which provides an important clue as to the extent to which monolingual domain corpora can be used in the present framework.

4 Experiments and evaluations

4.1 Experimental setup

4.1.1 Terminological dictionaries

For evaluation, we used five terminological dictionaries of computer science (henceforth COM) [5], economics (ECN) [6], law (LAW) [7], physics (PHY) [8] and psychology (PSY) [9]. These terminological dictionaries contain Japanese-English term pairs. The number of terms are listed in Table 1.

Table 1: Terminological dictionaries

Domain	# terms	Domain	# terms
COM	16,259	PHY	11,081
ECN	9,210	PSY	7,026
LAW	10,020		

Table 2 lists the number of terms by length (by the number of constituent units). The number of Japanese-English term pairs of which the number of constituent units is the same for Japanese and English is also listed (CP). There are a few terms

[3] http://developer.yahoo.co.jp/webapi/search/premium.html
[4] http://groonga.org/

[5] Aiso, H. 1993. *Joho Syori Daijiten*. Ohm, Tokyo.
[6] Yuhikaku, 1986. *Keizai Yougo Jiten*. Yuhikaku, Tokyo.
[7] Ozaki, T. 2003. *A Dictionary of English Legal Terminology*. Jiyukokuminsha, Tokyo.
[8] MeXT, 1990. *Japanese Scientific Terms: Physics*. Baifukan, Tokyo.
[9] MeXT, 1986. *Japanese Scientific Terms: Psychology*. Gakujutsu Shinkokai, Tokyo.

with no constituent units; they were produced because the POS-taggers mistakenly judged the constituent units to be functional elements rather than content words. Those listed in the rows "CP" with more than two constituent units are the sources of the head-modifier bipartite graph.

Table 2: Number of terms by length

		Number of constituent units				
		0	1	2	3	4+
COM	JA	47	4678	7849	3030	655
	EN	13	2522	8651	4055	1018
	CP	0	1857	6153	1893	247
ECN	JA	1	1970	5093	1641	415
	EN	6	1152	4883	2104	975
	CP	0	893	3725	928	153
LAW	JA	5	3738	4312	1320	645
	EN	31	3525	3954	1402	1108
	CP	0	2330	2381	416	93
PHY	JA	13	3480	6014	1401	173
	EN	7	2153	6453	2137	331
	CP	0	1703	4471	786	51
PSY	JA	21	2451	3633	806	115
	EN	17	2180	3737	904	188
	CP	0	1567	2723	367	31

4.1.2 Collecting web documents

The web documents for these domains were collected from November to December 2012. In collecting the web documents, the following domain keywords were used: 計算機科学 (keisanki kagaku) and "computer science" for COM, 経済学 (keizaigaku) and "economics" for ECN, 法学 (hougaku) and "law" for LAW, 物理学 (butsurigaku) and "physics" for PHY, and 心理学 (shinrigaku) and "psychology" for PSY. Table 3 shows the number of pages obtained from the web search. The "Japanese" and "English" columns show the number of pages obtained by using Japanese and English terms, respectively. Note that the number of Japanese web pages collected for COM is much smaller than its English counterpart, while in the other five domains they are more balanced.

We randomly selected 200 web pages for each domain, without distinguishing between English and Japanese pages, and checked the relevance of the pages to the domain. Table 4 shows the number of pages clearly relevant to the domain in question (official Web sites of relevant research institutes or departments of universities, international conference sites, terminological lists on the web, QR sites specific to the domain, forums dealing with relevant topics, blogs and essays written by domain experts) among the 200 pages for each domain. All in all, around 70 per cent of the collected web documents were relevant to the domain.

Table 3: Number of collected web pages

Domain	Japanese	English	Total
COM	4508	65440	69948
ECN	42556	58802	101358
LAW	30857	63804	94661
PHY	37905	67989	105894
PSY	29556	40539	70095

Table 4: Ratio of web pages relevant to the domain

Dom.	Relevant	Dom.	Relevant
COM	135 (67.5%)	PHY	160 (80.0%)
ECN	141 (70.5%)	PSY	137 (68.5%)
LAW	135 (67.5%)		

4.2 Generating term candidate pairs

Table 5 shows the basic statistics of the initial head-modifier bipartite graphs (created from steps 1–4 in section 3.2), in which "mods" stands for modifiers, "# comp" shows the number of connected components, "maxcmp" shows the number of vertices in the maximum component, "2nd cmp" shows the number of vertices of the second largest component (other headers should be obvious). As in many real-world networks, these initial graphs consist of one giant component and a number of small components (Newman, 2003; Newman, 2010). Using these initial graphs for generating potential MWT candidates would be unrealistic; a terminology of a domain cannot reasonably contain terms in the order of millions.

Table 6 shows the statistics of head-modifier graphs generated by removing bridges and applying the Kernighan-Lin algorithm. Essentially, the largest components in the initial graphs were partitioned into smaller components with similar sizes, while many previously connected vertices became isolated vertices. As a result, the number of po-

Table 5: Initial head-modifier bipartite graphs

Dom	# edges	# vertices	# heads	# mods	# comp	maxcmp	2nd cmp	# possible edges
COM	9,270	2,925	724	2,201	32	2,855	4	1,490,936
ECN	5,135	2,522	781	1,741	60	2,382	6	1,189,129
LAW	2,334	1,667	706	961	88	1,463	6	519,217
PHY	5,666	2,526	668	1,858	50	2,396	8	1,088,374
PSY	2,996	1,850	499	1,351	47	1,743	4	577,672

Table 6: Partitioned head-modifier bipartite graphs

Dom	# edges	# vertices	# heads	# mods	# comp	maxcmp	2nd cmp	# possible edges
COM	476	1,788	281	1,507	18	112	112	26,002
ECN	485	1,199	284	915	10	149	149	31,746
LAW	358	676	186	490	5	169	168	22,420
PHY	572	1,233	241	992	9	154	154	29,633
PSY	495	784	152	632	6	195	194	23,316

tential head-modifier candidates was reduced to the order of tens of thousands. Given the size of the original terminological dictionaries as well as many existing terminological dictionaries, this size seems reasonable.

4.3 Quantitative evaluations

Table 7 shows the number of candidate term pairs after validation (those pairs of which both Japanese and English candidates were validated at least once against the collected web documents were identified as candidate pairs).

Table 7: Number of term candidate pairs

Dom	# candidates	Jaccard > 0
COM	960	242
ECN	4,134	694
LAW	1,828	133
PHY	1,869	389
PSY	1,559	421

We manually evaluated (i) 100 top candidate pairs according to the Jaccard coefficient value, (ii) 100 top candidate pairs as calculated by the sum of Japanese and English hits, and (iii) 100 randomly chosen candidate pairs whose Jaccard coefficient was zero. They were evaluated from two points of view: according to (a) whether the Japanese and English matched, and (b) whether the Japanese candidate could be regarded as a term in the domain in question. For (b), we took into account cases in which the candidate was not in itself a term but could be a part of a longer term. Evaluation was carried out by two people; the results of the first evaluator were cross-checked by the other[10].

The results are listed in Table 8. The Jaccard coefficient gave the highest performance both in terms of bilingual correspondence (pairing) and in terms of validity to the domain. This indicates that the co-occurrence of SLT and TLT in the same document provides strong evidence for a pair being both a valid pair as well as valid terms. The low performance of law was due to the fact that the terminology of law we used contained many verbal expressions, which led to CU level mismatches (see section 4.4).

Unfortunately, the number of candidate pairs with a non-zero Jaccard coefficient was limited, as indicated in Table 7. However, it can be observed that the number of hits is also useful as evidence. In the present experiment we only used the sum of Japanese and English hits; we may be able to obtain more efficient information by taking into account the balance between the hits in the two languages.

Lastly, there are still relevant terms among the 100 randomly selected candidates, though the ratio of correct term pairs is much lower. To take full advantage of the proposed method, further filtering

[10] We did not carry out independent evaluations by multiple evaluators, as in real-world situations cross-checking is the more common method.

of the relevant terms from this range of candidates will be necessary.

Table 8: Results of evaluations

Jaccard			
Dom	pairing	term	partial term
COM	89	51	15
ECN	91	60	18
LAW	78	47	28
PHY	98	67	25
PSY	99	72	16
Hits			
Dom	pairing	term	partial term
COM	61	28	17
ECN	56	37	16
LAW	49	31	15
PHY	49	30	9
PSY	71	73	1
Random			
Dom	pairing	term	partial term
COM	30	15	2
ECN	24	23	11
LAW	32	12	6
PHY	37	17	13
PSY	45	37	12

4.4 Diagnosis

Changes in algorithms and parameter settings of the method, such as the bipartite graph partition algorithm or the selection of domain keywords for collecting web documents (note the substantially smaller number of Japanese computer science documents), will affect the behaviour of the system. In addition, it may be useful to make further use of the information which can be derived from the graph, such as the *degree* of vertices. These need to be examined systematically to optimise the performance vis-à-vis the nature of terminologies, which will be our future task in methodological front.

In addition, some general error patterns were observed upon closer qualitative observation:

- Especially for terms in the domain of law, errors arising from the mistreatment of postpositions and delimiting symbols in Japanese and prepositions in English were observed (e.g. the output "消費の行動" ("behaviour of consumption") is not a valid MWT, but "消費行動" ("consumption behaviour") without the postposision "の" ("of") could be. This problem can be solved by introducing MWT patterns or rules to restrict valid MWT forms to filter out candidates with invalid patterns.

- "Partial" terms were often generated and validated which in themselves are not valid MWTs but constitute a part of certain longer MWTs. This problem arises from the limitation of our method in which we only generate term candidates with two constituent elements. This shortcoming can be overcome by detecting maximum MWT patterns in the web documents. Rich accumulation of research in pattern-based MWT extraction can be directly relevant for this purpose (Ananiadou, 1994; Daille et al., 1994; Justeson and Katz, 1995; Nakagawa, 2000; Takeuchi et al., 2004).

- Some candidates which were judged as non-terms consist of two CUs which both represent generic concepts. To avoide this type of error, it will be useful to make use of the weight of vertices in the ogirinal graph, as well as using ontological information or introducing the idea of "stop words."

- Some errors arising from incorrect CU level pairing were observed as well. These can be avoided, at least partially, by introducing dictionary-based pairing of source language and target language CUs.

5 Conclusions and outlook

We proposed a way of augmenting bilingual terminologies by using a "generate and validate" method, taking advantage of the characteristics of terms and terminologies. The results of our experiments indicate that the method will be useful for collecting term candidate pairs to be included in existing terminological dictionaries.

The system which carries out this task is fully operational, although there is an uncertainty as to the free availability of the search engine api in future. For our system to be used in the real world in a Japanese-English setting, it should be complemented by methods which can detect and collect Japanese borrowed terms written in *katakana*, as they tend to be used for introducing new singular terms (Kageura, 2012). For this task, the "collect and validate" framework proposed by (Sato, 2010) would be useful, even though it was developed for collecting proper names.

In the next stage, we will evaluate the usefulness of the system *in vivo* rather than *in vitro*, in cooperation with dictionary companies, academic societies managing terminologies, and document management divisions of companies. A Japanese dictionary company has already expressed interest in trying our system in the process of revising some of its dictionaries.

References

Ananiadou S. 1994. A methodology for automatic term recognition. *COLING 1994*, 1034–1038.

Bernhard D. 2006. Multilingual term extraction from domain-specific corpora using morphological structure. *EACL 2006*, 171–174.

Chiao Y-C. and Zweigenbaum P. 2002. Looking for candidate translational equivalents in specialized, comparable corpora. *COLING 2002*, 1208–1212.

Cerbah F. 2000. Exogeneous and endogeneous approaches to semantic categorization of unknown technical terms. *COLING 2000*, 145–151.

Dagan I. and Church K. 1997. Termight: Coordinating humans and machines in bilingual terminology acquisition. *Machine Translation*, 12:89–107.

Daille B. et al. 1994. Towards automatic extraction of monolingual and bilingual terminology. *COLING 1994*, 515–521.

Daille B. and Morin E. 2008. Effective compositional model for lexical alignment. *IJCNLP 2008*, 95–102.

Daille B. and Morin E. 2012. Revising the compositional method for terminology acquisition from comparable corpora. *COLING 2012*, 1797–1810.

Fung P. and McKeown K. 1997. Finding terminology translations from non-parallel corpora. *5th Workshop on Very Large Corpora*, 192–202.

Gaussier E. 1998. Flow network models for word alignment and terminology extraction from bilingual corpora. *COLING/ACL 1998*, 444–450.

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

Kageura K. 2012. *The Quantitative Analysis of the Dynamics and Structure of Terminologies*. John Benjamins, Amsterdam.

Kernighan B. W. and Lin S. 1970. An efficient heuristic procedure for partitioning graphs. *Bell Systems Technical Journal*, 49:291–307.

Kwong O. Y. et al. 2004. Alignment and extraction of bilingual legal terminology from context profiles *Terminology*, 10(1):81–99.

Laroche A. and Langlais P. 2010. Revisiting context-based projection methods for term-translation spotting in comparable corpora. *COLING 2010*, 617–625.

Lefever E. et al. 2009. Language-independent bilingual terminology extraction from a multilingual parallel corpus. *EACL 2009*, 496–504.

Li B. and Gaussier E. 2010. Improving corpus comparability for bilingual lexicon extraction from comparable corpora. *COLING 2010*, 644–652.

Morin E. et al. 2010. Brains, not brawn: the use of "smart" comparable corpora in bilingual terminology mining. *ACM Transactions on Speech and Language Processing*, 7(1):Article 1.

Nakagawa H. 2000. Automatic term recognition based on statistics of compound nouns. *Terminology*, 6(2):195–210.

Newman M. E. J. 2003. The structure and function of complex networks. *SIAM Review*, 45(2):167–256.

Newman M. E. J. 2010. *Networks: An Introduction*. Oxford University Press, Oxford.

Robitaille X., et al. 2006. Compiling French-Japanese terminologies from the web. *EACL 2006*, 225–232.

Sager J. C. 1990. *A Practical Course in Terminology Processing*. John Benjamins, Amsterdam.

Sato S. 2010. Non-productive machine transliteration. *RIAO 2010*, 16–19.

Takeuchi K. et al. 2004. Construction of grammar-based term extraction model for Japanese. *Computerm 2004*, 91–94.

Tonoike M. et al. 2005. Effect of domain-specific corpus in compositional translation estimation for technical terms. *IJCNLP 2005*, 116–121.

Toutanova K. et al. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. *HLT-NAACL 2003*, 252–259.

TTC – Terminology Extraction, Translation Tools and Comparable Corpora. 2010–2012. `http://www.ttc-project.eu/`

Utsuro T. et al. 2007. Compiling bilingual lexicon for technical terms using the web. *ICKS 2007*, 27–34.

Van der Eijk P. 1993. Automating the acquisition of bilingual terminology. *EACL 1993*, 113–119.