Automatic Thesaurus Generation through Multiple Filtering

Kyo KAGEURA[†], Keita TSUJI[‡], and Akiko, N. AIZAWA[†]
[†]National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430 Japan
E-Mail: {kyo,akiko}@nii.ac.jp
[‡]Graduate School of Education, University of Tokyo,
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113 Japan
E-Mail: i34188@m-unix.cc.u-tokyo.ac.jp

Abstract

In this paper, we propose a method of generating bilingual keyword clusters or thesauri from parallel or comparable bilingual corpora. The method combines morphological and lexical processing, bilingual word alignment, and graph-theoretic cluster generation. An experiment shows that the method is promising.

1 Introduction

In this paper, we propose a method of automatic bilingual thesaurus generation by a combination of methods or multiple filtering. The procedure consists of three modules: (i) a morphological and lexical processing module, (ii) a translation pair extraction module, and (iii) a cluster generation module. The method takes parallel or comparable corpora as input, and produces as output bilingual keyword clusters with a reasonable computational cost.

Our aim is to construct domain-oriented bilingual thesauri, which are much in need both for cross-language IR and for technical translators. We assume domain-dependent parallel or comparable corpora as a source of information, which are abundant in case of Japanese and English.

The techniques used in each module are reasonably well developed, including statistical word alignment methods. However, there remain at least three problems: (i) ambiguity of multiple hapax combinations in an alignment, which cannot be resolved by purely statistical methods, (ii) syntagmatic unit mismatches, especially in such cases as English and Japanese, and (iii) difficulty in final cleaning-up[1] .

In this paper, we show that the proper combination of the above modules can be useful especially for resolving the cleaning-up problem and can produce good results in bilingual cluster or thesaurus generation.

2 Method

The procedure for thesaurus generation consists of the following three main modules.
(1) Morphological and lexical processing module: keyword units[2] for English and Japanese are extracted separately.
(2) Translation pair extraction module: statistical weighting is applied to a corpus which has been through the morphological and lexical processing module. The aim of this stage is not to determine unique translation pairs, but to restrict translation candidates to a reasonable extent.
(3) Cleaning and cluster generation module: a bilingual keyword graph is constructed on the basis of the pairs extracted at translation pair extraction module, and a graph-theoretic method is applied to the keyword graph, to generate proper keyword clusters by removing erroneous links.

[1] If we want to obtain a clean lexicon, minor translation variations tend to be omitted, while many errors would be included if we want to retain minor variations.

[2] The word 'keyword' implies words that are important with respect to documents or domains. In this paper, we use the word for convenience, roughly in the same sense as "content-bearing words". If necessary, a module of keyword or term weighting (e.g. Frantzi & Ananiadou 1995; Nakagawa & Mori 1998) can be incorporated easily.

2.1 Morphological & lexical processing

At this stage, basic lexical units or keyword candidates are extracted. We separately extract minimum or shortest units and maximum or longest complex units as syntagmatic units for keyword candidates. So two outputs are produced from this module, i.e. a bilingual keyword corpora of minimum units and another of maximum units.

The processing proceeds as follows:

(a) Morphological analysis

First, the corpus is morphologically analysed and POS-tagged. Currently, JUMAN3.5 (Kurohashi & Nagao 1998) is used for Japanese and LT_POS/LT_CHUNK (Mikheev 1996) is used for English.

(b1) Extraction of minimum units

Minimum units in English are simply defined as non-functional simple words extracted from the output of LT_POS. Minimum meaningful units in Japanese are defined as:

```
C_Prefix* (C_Adv|C_Adj|N) C_Suffix*
```

where `C_` indicates that the unit should consist of either Chinese characters or Katakana[3] .

(b2) Extraction of maximum units

Maximum complex units for English are the units extracted by LT_CHUNK, with some ad-hoc modifications.

Maximum complex units for Japanese are defined by the following basic pattern,

```
^C_Adj* (C_Affix|C_Adv|C_Adj|N)+
```

where `^C` means that the unit should begin with either Chinese character or Katakana. The pattern remains deliberately coarse, to absorb errors by JUMAN. Coarse patterns with simple character type restrictions produce better results than grammatically well-defined syntagmatic patterns. A separate stop word list for affixes is also prepared together with an exceptional treatment routine, to make the Japanese units better correspond to English units[4] .

After these processes, two corpora, one consisting of minimum units and the other of maximum units, are created.

Intermediate constituent units are not extracted, because their inter-lingual unit correspondence is less reliable. Also, many important intermediate units of longer complex units appear themselves as an independent complex unit in a large domain-specific corpus, and, even if they do not, intermediate units can be extracted on the basis of minimum and maximum translation pairs if necessary.

2.2 Extraction of translation candidates

The module for extracting translation candidate pairs consists of statistical weighting and postprocessing. These are applied to the data of minimum units and maximum units separately. After that, the two data are merged to make input for the cluster generation module.

(a) Statistical weighting

Many methods of extracting lexical translation pairs have been proposed (Daille, Gaussier & Langé 1994; Eijk 1993; Fung 1995; Gale & Church 1991; Hiemstra 1996; Hull 1998; Kupiec 1993; Melamed 1996; Smadja, McKeown & Hatzivassiloglou 1996). Though it is difficult to evaluate the performance of existing methods as they use different corpora for evaluation[5] , the performance does not seem to be radically different. We adopted log-likelihood ratio (Danning 1993), which gave the best performance among crude non-iterative methods in our test experiments[6] .

(b) Postprocessing filter

As the output of statistical weighting is simply a weighted list of all English and Japanese co-occurring pairs, it is necessary to restrict translation candidates so that they can be effectively used in the graph-theoretic cluster generation module. In addition to restricting possible translation pairs, it is necessary to determine unique translation pairs for hapax legomena. We use both macro- and micro-filtering heuristics to restrict translation candidates.

---

[3] In addition, we have made a few ad-hoc rules to screen out some consistent errors produced by the morphological analysers.

[4] For instance, the Japanese suffix '用' is eliminated because it corresponds in most cases to the English word 'for', which tends to be excluded from chunks made by LT_CHUNK.

[5] A common testbed exists for French-English alignment (Veronis 1996-99) but not for Japanese-English.

[6] At the time of writing this paper, we have finished a preliminary comparative experiments of various methods, among which the method proposed by Melamed (1996) gave by far the best result. We are thus planning to replace this module with the method proposed by Melamed (1996).

Two macro heuristics, applied to the overall list of pairs, are defined, i.e. (i) a proper translation should have a statistical score higher than the threshold X_S, and (ii) a keyword should have maximally X_C translations or $X_P \times$ *token frequency* when the frequency is less than X_C.

Micro heuristics uses the information within each alignment; we assume that a keyword in one language only has one translation within an alignment[7] . Selecting unique pairs in each alignment is achieved by recursively taking a pair with the highest score within an alignment, each time deleting other pairs which have the same English or Japanese elements[8] .

After this process, the data of minimum units and maximum units are merged, which constitutes input for the next stage.

2.3 Graph-theoretic cluster generation

Up to this stage, the cooccurrence information used to extract pairs has the depth of only one. In order to eliminate erroneous translations, we re-organise the extracted pairs into graph and use multi-level topological information by applying the graph-theoretic method. For explanation, let us assume that we obtained the data in Table 1 from the previous module as an input to this module.

Firstly, the *initial keyword graph* is constructed, where each node represents an English or Japanese keyword, and a link or edge represents the pairwise relation of corresponding keywords. We define the *capacity* or strength of a link by the frequency of occurrence of the pair in the corpus, i.e. the number of alignments in which the pair occurs[9] . Figure 1 shows the initial keyword graph made from the data in Table 1. The task now is to *detect and remove erroneous links* to generate independent graphs or clusters consisting of proper translation pairs[10] .

Japanese keywords	English keywords	frequency
キーワード	information retrieval	1
キーワード	keyword	39
テキスト検索	information retrieval	1
テキスト検索	text retrieval	6
テキスト検索	text search	3
検索指示語	keyword	1
広域情報検索	information retrieval	1
情報検索	information gathering	4
情報検索	information retreival	1
情報検索	information retrieval	320
情報検索	information search	5
情報収集	information gathering	6
情報収集	information retrieval	1
文献検索	bibliographic search	1
文献検索	document retrieval	11
文書検索	document retrieval	19
文書検索	text retrieval	1

Table 1. Input example for cluster generation

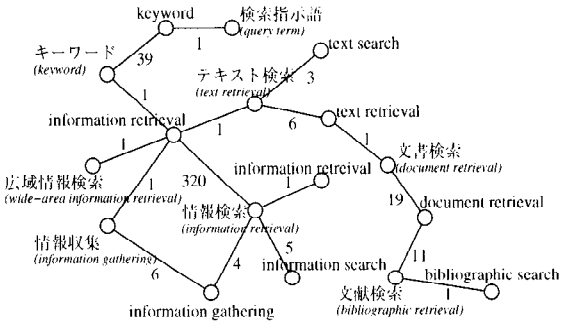

Figure 1. Example of initial keyword graph

The detection algorithm is based on the simple principle that sets of links, which decompose a connected keyword cluster into disjoint subclusters when they are removed from the original cluster, are candidates for improper translations. In graph theory, such a link set is called an edge cut and the edge cut with the minimal total capacity is called a *minimum edge cut*. A minimum edge cut does not necessarily imply a single translation error. An efficient al-

[7] This is not true for longer alignment units such as full texts. However, this will apply to parallel titles and abstracts which are readily available. Many lexical alignment methods starting from sentence-level alignments assume this or some variations of this.

[8] Many maximum unit pairs in fact have the same score. We used the arithmetic mean of the constituent minimum units to resolve alignment ambiguity.

[9] The score of likelihood ratio is a possible alternative for link capacity, but the result of a preliminary experiment was no better. In addition, after selecting pairs by threshold, whether a pair constitutes a proper translation or not is not a matter of weight, because threshold setting implies that all pairs above that are regarded as correct. So we adopt simple frequency as the link capacity. However, we notice a lack of affinity between the statistical alignment method we used here and the definition of link capacity, which is currently under examination and will be improved by renewing the alignment module.

[10] This approach is radically different from statistically oriented word clustering (Deerwester et. al 1990; Finch 1993; Grefenstette 1994; Schütze & Pedersen 1997; Strzalkowski 1994); this is why we use the word 'cluster generation' instead of 'clustering'.

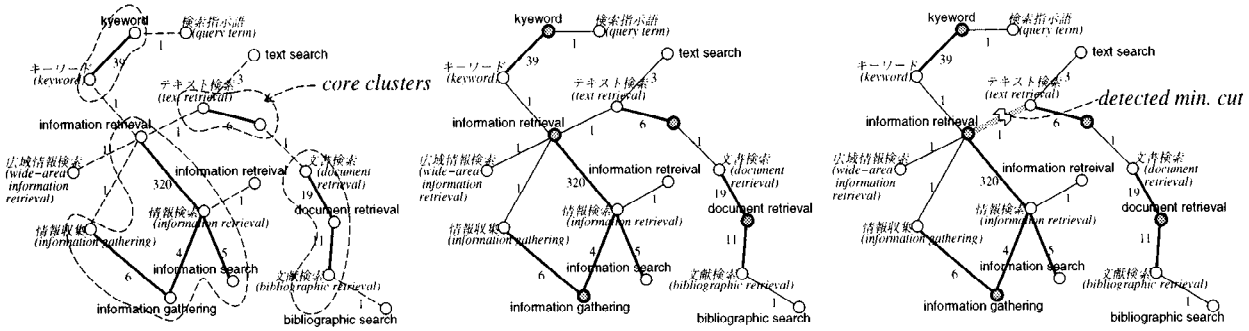

Figure 2. Steps of graph-theoretic cluster generation

gorithm exists for minimum edge cut detection (Nagamochi 1993).

Our procedure first checks links that should not be eliminated, using the conditions: (i) the frequency is no less than N_α, (ii) the Japanese and English notations are identical, or (iii) either of the Japanese or English expressions have only one corresponding translation (Figure 2 (a); it is assumed that $N_\alpha = N_\beta = N\epsilon = 3$). Secondly, *core* keywords whose frequency is no less than N_β are checked (Figure 2 (b)). This is used for the restriction that each cluster should include at least one *core* keyword. Lastly, *edge cuts* with a total capacity of less than N_ϵ are detected and removed (Figure 2 (c)). This procedure is repeated recursively until no further application is possible. Figure 3 shows the state after these steps are applied.

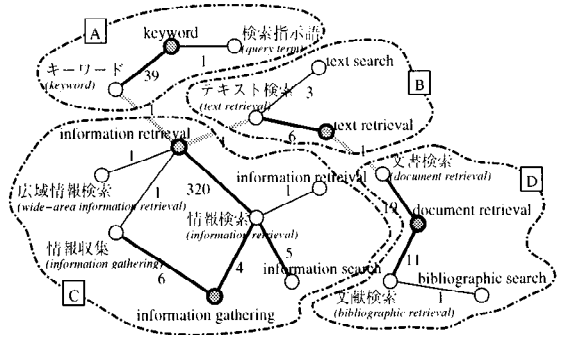

Figure 3. Generated clusters

3 Experiment

3.1 Settings and procedures

We applied the method to Japanese and English bilingual parallel corpus consisting of 25534 title pairs in the field of computer science. Table 2 shows the basic quantitative information after morphological and lexical filtering was applied.

Minimum units				
Japanese	Token:	178091	Type:	14938
English	Token:	154554	Type:	12634
Maximum units				
Japanese	Token:	89742	Type:	38813
English	Token:	80018	Type:	41693

Table 2. Basic quantity of the data

In the pair extraction module, the threshold X_S was set to 10^{11} . The parameter X_C was set to 10 and X_P to 0.5. As a result, 28905 translation candidate pairs were obtained, with 24855 Japanese and 23430 English keywords. Of these, 20071 pairs occurred only once and 3581 only twice. The most frequent pair occurred 3196 times in the corpus. 8242 (28.5%) were minimum unit pairs, and 20663 (71.5%) were maximum unit pairs.

Table 3 shows the number of keywords which had N translations. On average, a Japanese keyword had 1.16 English translations, while an English keyword had 1.23 Japanese translations.

N	Jap.	Eng.	N	Jap.	En.
1	21796	19778	5	62	157
2	2409	2693	6	10	59
3	412	437	7	7	17
4	159	285	8	0	4

Table 3. Number of translations

[11] This is purely heuristic. Minimum units and maximum units are given different scores. But only 3 pairs below this threshold were proper translation pairs in 100 random samples of minimum unit pairs, and 5 in 100 samples of maximum units.

Evaluating recall and precision on the basis of 100 randomly selected title pairs, which consisted of 778 keyword token pairs, the precision tokenwise was 84.06% (654 correct translations) and the recall was 87.08% (654 of 751 correct pairs). Typewise precision was 81.65% (543 correct of 665 pairs).

The initial keyword graph generated from these 28905 translation candidates consisted of 19527 independent subgraphs, with the largest cluster containing 2701 pairs (i.e. 9.3% of all the pairs). The cluster generation method was applied with parameters $N_\alpha = 4$, $N\epsilon = 10$ and $N_\beta = 1$[12] . As a result, 893 translation pairs were removed, and 20357 bilingual clusters were generated. The maximum cluster now contained only 64 pairs. Table 4 shows the number of clusters by size given by number of pairs.

size	no. of clusters	size	no. of clusters
1	16693	5-9	322
2	2354	10-19	52
3	504	20-64	22
4	410		

Table 4. Number of clusters by size

3.2 Overall evaluation

The result was manually evaluated from two points of view, i.e. consistency of clusters and correctness of link removal[13] .

(1) To check the internal consistency, clusters were classified into three groups by size, and were separately evaluated. 2000 'small' clusters, consisting of only one pair, were randomly sampled and evaluated as 'correct' (c), 'more or less correct' (m) or 'wrong' (w). 400 medium size clusters consisting of 2-9 pairs and all the 74 large clusters consisting of 10 or more pairs were evaluated as 'consistent' (c: consisting only of closely related keywords), 'mostly consistent' (m: consisting mostly of related keywords), 'hybrid' (h: consisting of two or more different keyword groups: h) or 'bad' (w). Table 5 shows the result of the evaluation. The general performance is very good, with more or less 80% of the clusters being meaningful.

[12] This is again determined heuristically. For an examination of the effect of parameters, see Aizawa & Kageura (to appear).

[13] The evaluation was done by the first author. Currently no cross-checking has been carried out.

For small clusters, the performance was separately evaluated for minimum and maximum unit pairs. Note that the ratio of maximum unit pairs is comparatively higher in the small cluster than the overall average. Most pairs evaluated as partially correct, as well as some wrong pairs, suffered from mismatch of the syntagmatic units.

	c	m	w	total
Small	1389	370	241	2000
	(69.5%)	(18.5%)	(12.1%)	(100%)
minimum	288	26	69	383
	(75.2%)	(6.8%)	(18.0%)	19.2%
maximum	1101	344	172	1617
	(68.1%)	(21.3%)	(10.6%)	80.9%
	c	m	h	w
Medium	116	148	32	104
	(29.0%)	(37.0%)	(8.0%)	(26.0%)
Large	8	18	43	5
	(10.8%)	(24.3%)	(58.1%)	(6.8%)

Table 5. Evaluation of internal consistency

73% of the medium sized clusters were 'correct', 'mostly correct' or 'hybrid'. Among the 'mostly correct' and 'hybrid' clusters, 97 (91 and 6 respectively) were mainly caused by the mismatch of the units. For instance, in the case: { 最適, 最適化, 最適な, *optimization, optimal, optimisation, optimum, network optimization* }, the last English keyword has the excess unit 'network'. Other 'mostly correct' and 'hybrid' clusters were due to the problem of corpus frequencies.

Among the large clusters, more than half were 'hybrid'[14] . Among the 'mostly correct' and 'hybrid' large clusters, only 8 (3+5) were due to unit mismatch, while 53 (15+38) were due to quantitative factors. This shows a striking contrast to the medium sized clusters. Large hybrid clusters tended to include many common word pairs which occur frequently. For instance, in the largest cluster, 'システム system' (3196), '開発 development' (1097), '設計 design' (1073), and '環境 environment' (890) are included due to indirect associations. The following are two examples of hybrid clusters, whose hybridness comes from quantitative factors and unit mismatches respectively:

Example 1: 概要/全体/要約/サマリ/overview/outline/summary/summarization/overall

[14] And most of the sub-clusters in these hybrid clusters are 'mostly correct'.

Example 2: パターン/パタン/パターン認識/照合/パターンマッチング/パターン照合法/パターン照合/pattern/patterns/patten/patterm matching

In the first case, the 'overall' group and the 'summary' group are mixed up. In the second case, the mismatch of syntagmatic units is caused by borrowed words. In fact, many errors caused by the mismatch of syntagmatic units involve borrowed words written in Katakana.

(2) To look at the performance of graph-theoretic cluster generation, we examined the removed pairs from two points of view, i.e. the correctness of link removal and the internal consistency of clusters generated by link removal. For the former, we introduced three categories for evaluation: mismatched pairs correctly removed (c), proper translation pairs wrongly removed (w), and pairs of related meaning removed (p). The consistency of newly generated clusters were evaluated in the same manner as above.

	c	p	w	total
cc	90 (10.1)	53 (5.9)	39 (4.4)	182 (20.4)
cm	148 (16.6)	56 (6.6)	32 (3.6)	236 (26.4)
ch	96 (10.8)	20 (2.2)	6 (0.7)	122 (13.7)
mm	44 (4.9)	29 (3.3)	30 (3.4)	103 (11.5)
mh	52 (5.8)	13 (1.5)	5 (0.6)	70 (7.8)
hh	30 (3.4)	3 (0.3)	3 (0.3)	36 (4.0)
xc	42 (4.7)	9 (1.0)	9 (1.0)	60 (6.7)
xm	28 (3.1)	8 (0.9)	20 (2.2)	56 (6.3)
xh	8 (0.9)	2 (0.3)	5 (0.6)	15 (1.7)
xx	4 (0.5)	1 (0.1)	8 (0.9)	13 (1.5)
all	542 (60.7)	194 (21.7)	157 (17.6)	893 (100)

Table 6. Evaluation of removed links

Table 6 shows the result of evaluation of all the 893 removed pairs. 'c' 'p' and 'w' in the top row indicate types of removed links, and 'cc', 'cm' etc. in the leftmost column indicate internal consistencies of two clusters generated by link removal. A total of 157 (17.6%) of the removed links were correct links wrongly removed, but among them, 115 links did not produce 'bad' clusters. If we consider them to be tolerable, only 42 removals (4.7%) were fatal errors.

By examining the removed links, we found that the links removed at the higher edge capacity included more wrongly removed pairs. For instance, among 142 edges removed at capacity 4 (which is the maximum deletable value set by N_α), 41 or 28.9% were wrongly removed correct translations, while among 288 links removed at capacity 1, only 15 or 5.2 % were correct translations.

4 Discussion

From the experiment, we have found some factors that affect performance.

(1) Many errors were produced at the stage of extracting keyword units, by syntagmatic mismatch. A substantial number of them involved Japanese Katakana keywords. Therefore, in addition to the general refinement of the morphological processing module, the performance will be improved if we use string proximity information to determine syntagmatic units[15] .

(2) We expect that some errors produced by statistical weighting and filtering could be removed by applying stemming and orthographic normalisations, which are not fully exploited in the current implementation. Looking back from the cluster generation stage, frequently occurring keywords tend to cause problems due to indirect associations. At the time of writing, we are radically changing the statistical alignment module based on Melamed (1996) and incorporating iterative alignment anchoring routine so that the method can be applied not only to titles but also to abstracts, etc. Used in conjunction with string proximity and stemming information, we might be able to retain minor variations properly.

(3) At the cluster generation stage, we observed that correct links tend to be wrongly removed for higher capacities of edge cut. In the current implementation, the parameter values remain the same for all the clusters. Performance will be improved by introducing a method of dynamically changing the parameter values according to the cluster size and the frequencies of their constituent pairs.

5 Conclusion

We have proposed a method of constructing bilingual thesauri automatically, from parallel or comparable corpora. The experiment showed that the performance is fairly good. We are currently improving the method further, along the lines discussed in the previous section. Further experiments are currently being carried out, using the data of narrower domains (e.g. artificial

[15] This can also be used for resolving hapax ambiguity.

intelligence) as well as abstracts instead of titles.

At the next stage, we are planning to evaluate the method from the point of view of performance of generated clusters in practical applications. We are currently planning to apply the generated clusters to query expansion and user navigation in cross-lingual IR, as well as to on-line dictionary lookup systems used as translation aids.

Acknowledgement

This research is a part of the research project "A Study on Ubiquitous Information Systems for Utilization of Highly Distributed Information Resources", funded by the Japan Society for the Promotion of Science.

References

[1] Aizawa, A. N. and Kageura, K. (to appear) "A graph-based approach to the automatic generation of multilingual keyword clusters." In: Bouligault, D., Jacquemin, C. and l'Homme, M-C. (eds.) *Recent Advances in Computational Terminology.* Amsterdam: John Benjamins.

[2] Dagan, I. and Church, K. (1994) "Termight: Identifying and translating technical terminology. " *Proc. of the Fourth ANLP.* p.34–40.

[3] Daille, B., Gaussier, E. and Langé, J. M. (1994) "Towards automatic extraction of monolingual and bilingual terminology." *COLING'94.* p. 515–521.

[4] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. and Harshman, R. (1990) "Indexing by latent semantic analysis." *JASIS.* 41(6), p. 391–407.

[5] Dunning, T. (1993) "Accurate methods for the statistics of surprise and coincidence." *Computational Linguistics.* 19(1), p. 61–74.

[6] Eijk, van der P. (1993) "Automating the acquisition of bilingual terminology." *Proc. of the 6th EACL.* p. 113–119.

[7] Finch, S. P. (1993) *Finding Structure in Language.* PhD Thesis. Edinbourgh: University of Edinbourgh.

[8] Frantzi, K. T. and Ananiadou, S. (1995) "Statistical measures for terminological extraction." *Proc. of 3rd Int'l Conf. on Statistical Analysis of Textual Data.* p. 297–308.

[9] Fung, P. (1995) "A pattern matching method for finding noun and proper noun translations from noisy parallel corpora." *Proc. of 33rd ACL.* p. 233 236.

[10] Gale, W. A. and Church, K. W. (1991) "Identifying word correspondences in parallel texts." *Proc. of DARPA Speech and Natural Language Workshop.* p. 152–157.

[11] Grefenstette, G. (1994) *Explorations in Automatic Thesaurus Discovery.* Boston: Kluwer Academic.

[12] Hiemstra, D. (1996) *Using Statistical Methods to Creat a Bilingual Dictionary.* MSc Thesis, Twente University.

[13] Hull, D. A. (1998) "A practical approach to terminology alignment." *Computerm'98.* p. 1–7.

[14] Kitamura, M. and Matsumoto, Y. (1997) "Automatic Extraction of Translation Patterns in Parallel Corpora." *Transactions of IPSJ.* 38(4), p. 727–735.

[15] Kupiec, J. (1993) "An algorithm for finding noun phrase correspondences in bilingual corpora." *Proc. of 31st ACL.* p.17–22.

[16] Kurohashi, S. and Nagao, M. (1998) *Japanese Morphological Analysis System Juman version 3.5 User's Manual.* Kyoto: Kyoto University.

[17] Melamed, I. D. (1996) "Automatic construction of clean broad-coverage translation lexicons." *2nd Conference of the Association for Machine Translation in the Americas.* p. 125–134.

[18] Mikheev, A. (1996) "Learning part-of-speech guessing rules from lexicon." *COLING'96,* p. 770–775.

[19] Nagamochi, H. (1993) "Minimum cut in a graph." In: Fujisige, S. (ed.) *Discrete Structure and Algorithms II* (Chapter 4). Tokyo: Kindaikagakusha.

[20] Nakagawa, H. and Mori, T. (1998) "Nested collocation and compound noun for term extraction." *Computerm'98.* p. 64–70.

[21] Schütze, H. and Pedersen, J.O. (1997) "A cooccurrence-based thesaurus and two applications to information retrieval." *Information Processing and Management.* 33(3), p.307–318.

[22] Smadja, F., McKeown, K. R. and Hatzivassiloglou, V. (1996) "Translating collocations for bilingual lexicons: A statistical approach." *Computational Linguistics.* 22(1), p. 1–38.

[23] Strzalkowski, T. (1994) "Building a lexical domain map from text corpora." *COLING'94,* p.604–610.

[24] Veronis, J. (1996-) "ARCADE: Evaluation of parallel text alignment systems." http://www.lpl.univ-aix.fr/projects/arcade/

[25] Yonezawa, K. and Matsumoto, Y. (1998) "Zoshinteki taiouzuke ni yoru taiyaku tekisuto kara no hon'yaku hyougen no cyusyutu." *Proc of the 4th Annual Meeting of the Association for NLP.* p. 576–579.