Japanese and English Cross-lingual Information Retrieval at DLUT

Seigo Tanimura[†], Masashi Suzuki[‡], Hiroshi Nakagawa[†] and Tatsunori Mori[‡]

[†]Information Technology Center, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, JAPAN

[‡]Division of Electrical and Computer Engineering, Faculty of Engineering,
Yokohama National University
79-5 Tokiwadai, Hodogaya-ku, Yokohama-shi, Kanagawa, JAPAN

E-mail: tanimura@r.dl.itc.u-tokyo.ac.jp, masashi@forest.dnj.ynu.ac.jp,
nakagawa@r.dl.itc.u-tokyo.ac.jp, mori@forest.dnj.ynu.ac.jp

1 Abstract

We describe our cross-lingual retrieval system for cross-lingual information retrieval task of NTCIR2. The main part of our work is construction of bilingual dictionaries per academic field from a non-parallel bilingual corpus. EDICT is the start-point of constructing our bilingual dictionaries. We merge compound word translation extracted from a bilingual corpus to the dictionaries. Disambiguation of our bilingual dictionaries is performed using the bilingual corpus. Experimental evaluation of our bilingual dictionaries and document retrieval is depicted.

2 Introduction

A bilingual dictionary is a powerful information resource for cross-lingual information retrieval(CLIR). Translating a keyword by looking up a bilingual dictionary is generally much cheaper in the computational cost involved than translating a whole document by a machine translator. Nevertheless, translation of a keyword by a bilingual dictionary is not trivial. Keywords extracted from documents for indexing are sometimes compound words. A compound word of itself is usually not held in a bilingual dictionary. We are, however, not doomed to failure. Since a bilingual dictionary may hold a part of the words in a compound word, we can overcome the problem of translating a compound word by translating each of the words in a compound word individually, followed by constructing the possible translations of the compound word from the translated individual words. While a number of possible translation can be generated, only a part of them are correct. Hence disambiguation of the compound word translations from the constructed ones is crucial. Ambiguity in the translation of a compound word can be solved by building a bilingual dictionary per domain, assuming that a compound word translates into only one translation in a domain.

We describe our retrieval system for NTCIR2. The major part of our system is construction of bilingual dictionaries for query tranlsation. In section 3, we brief the work related to ours. The overview of our retrieval system is described in section 4. We then discuss the experimental evaluation of our system in section 6. The conclusion is given in section 7.

3 Related Work

It is generally desirable in cross-lingual document processing to choose a translation method with little deterioration of performance caused by translation. Collier et al. [Collier et al. 1998] compares the performance of translation by two types of methods, namely, dictionary lookup and machine translation, applied to aligning 1488 news articles written in Japanese to 6782 news articles written in English. A news article written in Japanese is translated into English by translating the keywords extracted from the article written in Japanese using a bilingual dictionary and translating the whole article written in Japanese using a machine translator. The news article translated into English is then matched against news articles written in English to seek the most similar one from them. Their results show that translation by dictionary lookup performs as well as by machine translator. Sakai et al. [Sakai et al. 1999] reports that while translation of a document to be searched generally outperforms translation of a query in CLIR, machine translation lacks of scalability against the size of documents. Considering the time consumed for machine translation and the number of documents we have to process, we have chosen an approach to translate a query by a bilingual dictionary.

The major problem of translating a query by looking up a bilingual dictionary is which kind of a dictionary should be adopted. A bilingual dictionary that can be adopted to CLIR is roughly classified into several types, described hereafter.

- A bilingual dictionary built from scratch

  Most of the methods to construct a bilingual dictionary from scratch adopt a parallel bilingual corpus, that is, a bilingual corpus consisting of documents and their translations. Documents, sentences of documents and sometimes words of sentences in a parallel corpus can be aligned to seek a pair of a document, sentence or word and its translation. We assume that the translation of a single or compound word in an alignment unit written in a language exists in the corresponding alignment unit written in the other language. A parallel corpus fairly satisfies this assumption. Kuipec et al. [Kupiec1993] extract the translation of a continuous noun phrase from a parallel corpus prealigned by sentences. Dagan and Church [Dagan and Church1994] translate a continuous technical term using a parallel bilingual corpus prealigned by words. Smadja et al. [Smadja et al. 1996] extract the translation of an interrupted collocation in addition to a continuous one from a parallel bilingual corpus prealigned by sentences. We cannot, however, simply apply the methods described insofar as the NTCIR2 test collection from which we retrieve documents is not a parallel corpus, that is, most of the documents written in Japanese have no corresponding documents translated into English, and vice versa. In addition, the fatal problem of constructing a bilingual dictionary from a parallel corpus is that a large parallel corpus written in languages with completely different characters or syntax rules is hardly available. Moreover, only a small amount of parallel corpora in limited domains are available at the best even in languages with similar characters or syntax rules due to a large amount of translation work involved.

  Some of the participants in the past NTCIR workshop also build a bilingual dictionary by methods using a parallel bilingual corpus. Chen et al. [Chen et al. 1999] generate a bilingual dictionary from keywords in a prealigned pair of documents written in Japanese and English. The keywords are the ones given by the authors of the documents. The first keyword in a document written in Japanese and the first keyword in the corresponding document written in English are extracted as the translation of each other, the second keywords in a Japanese document and the corresponding English document are extracted, and so forth. While their method is quite simple, it cannot be applied to documents with no keywords annotated by authors. Mitsuhiro et al. [Mitsuhiro et al. 1999] construct a bilingual dictionary with translation similarity from a parallel bilingual corpus based on an idea that the terms in aligned documents are similar. They report that the smaller size of a parallel corpus for building a bilingual dictionary results in the poorer precision with almost no change of recall, indicating increase of spurious translation in the bilingual dictionary.

  Transliteration is a promising method to construct a bilingual dictionary of the words borrowed from a foreign language including proper nouns from scratch without a parallel corpus. Since an existing bilingual dictionary available to us lacks of quite a few borrowed words, it is desirable to perform transliteration not only when we build a bilingual dictionary from scratch but also we adopt an existing bilingual dictionary. Transliteration is also superior to the methods using a parallel corpus in that we can apply transliteration to the words written in languages with somewhat different characters from each other. Knight and Graehl [Knight and Graehl1998] report their techinique developed to transliterate a word written in English into the corresponding word written in Japanese. They build a stocastic model of phonome translation using 8000 pairs of words in English and borrowed words in Japanese. Although they assume that an English word always translates into a borrowed words in Japanese, we often hit ourselves into a case where we translate an English word into not a borrowed word but a native word in Japanese. Fujii and Ishikawa [Fujii and Ishikawa1999] also construct a bilingual dictionary of borrowed words by transliterating a word in English into a borrowed word in Japanese. They avoid attempting to transliterate a word in English that translates into a native word in Japanese by first looking up two existing bilingual dictionaries, followed by looking up the word in the foreign word dictionary built by transliteration if the word to translate is not found in the existing dictionaries. Both of those methods transliterates an English word into a Japanese word because the number of phonomes in English is larger than that in Japanese. We adopt transliteration to construct a bilingual dictionary as well, except that we avoid transliterating a word in English that translates into a native word in Japanese by a measure of distance obtained from transliteration of an English word into a Japanese word.

- An existing bilingual dictionary

  EDICT [Breen1995] is a machine-readable Japanese-English dictionary consisting of about 64000 entries. In each of the entries, a word in Japanese is associated with several possible translations in English. As the actual concept of each of the possible translations differs from one another, translated words have to be disambiguated. Most of the past work to solve an ambiguous word are based on the assumption that the usage of a word in a sentence or a document is independent from the language of the word, and the usage depends on the concept of the word. The usage of a word can be computed by several methods. Syntactic characteristics can be adopted as the usage of a word by performing syntactic parse. Wehrli [Wehrli1998] translates an idiom, or a compund word consisting of a verb, using the syntactic structure of an idiom. Matsumoto et al. [Matsumoto et al. 1993] applies a similar method for Japanese-English translation, while Tzoukermann et al. [Tzoukermann et al. 1997] apply it to English-French translation. The promlem of those methods is that we cannot always determine the syntactic structure of a sentence or a compound word uniquely. As the larger bilingual corpus calls for the more number of possible syntactic structures, it is not suitable for a large bilingual corpus to adopt syntactic characteristics as the usage of a word.

  Another method to compute the usage of a word is adopting information of the words cooccurring around the word. Rapp [Rapp1995] disambiguates translation of 100 words in English to 100 words in German translated from the English words by hand using words cooccuring within 11 words from the English word and the German words in English and German non-parallel corpora. The measure of cooccurence is mutual information. Tanaka and Iwasaki [Tanaka and Iwasaki1996] apply Rapp's technique to disambiguate translation of 378 words using Japanese and English non-parallel corpora. They obtain the possible translations from EDICT. While Tanaka's method achieves outstanding performance of disambiguation, the computational cost involved in disambiguation is extremely expensive due to the large amount of context information per word to be disambiguated. Lin et al. [Lin et al. 1999] applies Tanaka's method, except that they count words cooccuring within 3 words from a word to be disambiguated.

  Fung [Fung1995] solves the problem of expensive computational cost by introducing context heterogeneity, that is, the number of distinct words cooccuring next to a word to be disambiguated standardized by the number of cooccurrence of the word, applied to disambiguation of translation between Chinese and English. Although her technique involves the context of an ambiguous word as Rapp, Tanaka and Lin do, context heterogeneity can be computed at an incredibly low computational cost because the amount of context information per ambiguous word is much smaller than that of the methods by Rapp and Tanaka. Although Fung disambiguate only 58 translations, we have to process a far more number of ambiguious words in practice than she does.

  Nakagawa [Nakagawa2000] proposes a technique to disambiguate the ambiguous words found in EDICT. His technique is unique in a couple of points. First, the number of occurrence of a word or words to be disambiguated is not involved in measuring the context of a word or words. He measures the context of a word or words to be disambiguated by only the numbers of the distinct words cooccuring directly prior and posterior to the word. He claims that his measure depends on not the size but the coverage of the bilingual corpus adopted for disambiguation, while mutual information and context heterogeneity depend on the size. This is suitable for adopting a bilingual corpus covering various domains. Second, the measured value of the context of a word is not directly applied to disambiguation. The words to be disambiguated are ranked by each of the measured value of the context, followed by standardizing the rank of each of the words by the number of the distinct words to be disambiguated. Then disambiguation is performed based on the idea that words in each of the languages appear at the same standerdized rank if the words share the identical concept. Experimental evaluation shows that his disambiguation method achieves almost the same performance for both non-parallel and parallel bilingual corpora.

4 Overview of Our System

Our system consists of three elements, briefed below.

- Construction of bilingual dictionaries

  This is the main part of our system. Using EDICT as our startpoint, we construct bilingual dictionaries per academic field from NTCIR1 test set collection [Kando et al. 1999]. The specific approach is described in section 5.

- Translation of keywords in a query

  Our search engine accepts a set of weighted keywords as a query. The engine searches either Japanese or English documents at a time, not both of them. The languages of a query and documents need to be the same. Hence if the language of a query differs from the language of the documents to be searched, we translate the keywords of a query by the bilingual dictionary covering the most similar academic field to the query in prior to search.

- Search engine

  We extract keywords weighted by TFIDF from each of the documents to be searched in advance. The similarity of a query and each of the documents to be searched is computed as the cosine of keyword vectors generated from the query and the document. The documents are then ranked in the descending order of the similarity. For mixed-language document retrieval, we search document sets in Japanese and English in parallel, followed by merging both of them to rank in the order of the similarity.

5 Construction of Bilingual Dictionaries

The source bilingual corpus of our bilingual dictionaries, namely NTCIR1 test set collection, is not a parallel corpus, making it difficult for us to build bilingual dictionaries from scratch. We thus adopt EDICT in order to build bilingual dictionaries. As described in section 3, the translations found in EDICT should be disambiguated so that a keyword is not translated into a word with different concept from the original keyword. Since NTCIR1 test set collection covers various academic domains and the number of documents per domain varies widely, we perform disambiguation by a technique based Nakagawa's method [Nakagawa2000], assuming that the number of the distinct words in a domain does not vary so drastically as the number of words in a domain does.

As the concept denoted by a word usually depends on the domain of the document where the word is written, we also have to take the domain of the word to be translated into account so that we can perform disambiguation properly. For example, an English word *architecture* should be translated into a Japanese word 建築 in the domain of architecture, while *architecture* should be translated into アーキテクチャ in the domain of computer science. Now the problem we must solve is how we should determine the domain of documents in which we can translate a word with no ambiguity. We assume that a word used in an academic field denotes only one concept, and documents in one society cover one academic field. Hence we build bilingual dictionaries per academic fields. Since more than one societies may cover the same academic field, the societies are cluserized into six groups. We construct bilingual dictionaries per those six groups of the societies.

Although compound words are frequently used as keywords, most of them are not found in EDICT. We attempt to solve this problem by translating a compound word that is not found in EDICT individially, followed by constructing the translation of the compound word from each of the individially translated words. As a number of possible translation of the compound word can be constructed, we again have to disambiguate them. In order to solve this problem, we assume that the documents written in Japanese and English convering the same academic field consist of words denoting an identical concept in both Japanese and English. We thus disambiguate translated compound words by extracting only those that appear in a corpus.

Since the words to be translated sometimes consist of borrowed proper nouns not found in EDICT, we adopt transliteration to extract the translation of borrowed proper nouns. As translation of a compound word is built by EDICT, translation of proper nouns extracted by transliteration should be merged into EDICT prior to constructing translation of a compound word. We thus filter out translation extracted by transliteration with an edit distance longer than a predetermined threshold, followed by merging the translation into EDICT.

The specific steps undertaken are as follows:

1. Clustering of documents in the source bilingual corpus

   For each of the societies covered by NICIR1 test set collection, construct a keyword vector where each of the keywords are weighted by its TFIDF value. Then clusterize the societies by the cosine similarity between the keyword vectors of two societies.

2. Transliteration

   Add translation extracted by tranliteration of an English word and an Japanese word written in Kanatana.

3. Disambiguation of the words in EDICT

   First measure the usage of a word to be disambiguated for each of the society groups and languages by the method described hereafter. Let $w = N_1 N_2 \ldots N_k \ldots N_K$ be a word in EDICT to be disambiguated, where N_k denotes a single word. For each of w, compute $Imp\,(w)$ using expression (1)

$$\begin{aligned} & Imp\,(w) \\ = & \prod_{k=1}^{K} \left((Pre\,(N_k) + 1) \cdot (Post\,(N_k) + 1) \right)^{\frac{1}{2K}} \end{aligned} \quad (1)$$

   where $Pre\,(N_k)$ and $Post\,(N_k)$ denote the numbers of the distinct words cooccuring directly prior and posterior to N_k in a bilingual corpus, respectively.

   Then rank the words to be disambiguated in the descending order of $Imp\,(w)$. The rank of a word is standardized by the number of distinct words to be disambiguated. Finally, disambiguate the possible translations $w_1, w_2, \ldots, w_n$ of a word w for each of the society groups by applying the following rule:

   Rule: Let $Ra\,(w)$ denote the standardized rank of w. If an inequality $|Ra\,(w) - Ra\,(w_i)| < |Ra\,(w) - Ra\,(w_j)|$ satisfies, then take w_i as the preferred translation of w over w_j.

   Hence $w_1, w_2, \ldots, w_n$ are ranked in the order of preference as the translation of w.

4. Construction and disambiguation of compound word translation

   First extract possible compound words from the source bilingual corpus by matching predetermined part-of-speech patterns. For each of the extracted compound words, the possible translations of the compound word are built by translating each of the words in the compound word using EDICT. We perform this process from Japanese to English and from English to Japanese. Then the translations are disambiguated by adopting the following rule:

   Rule: Let a compound word in Japanese, w_j have its possible translations in English, $W_e = \{w_{e1}, w_{e2}, \ldots\}$, where w_{e1}, w_{e2} and so forth are the possible translations. Let a compound word in Japanese, w_e have its possible translations in Japanese, $W_j = \{w_{j1}, w_{j2}, \ldots\}$, where w_{j1}, w_{j2} and so forth are the possible translations. If both $w_j \in W_j$ and $w_e \in W_e$ satisfy, then w_j and w_e are the translation of each other.

   If the disambiguated translations are still not unique, apply the same method as for disambiguation of the words in EDICT.

6 Experimental Evaluation

We evaluate our system experimentally, focusing at two points: the constructed bilingual dictionaries and document retrieval.

1. The constructed bilingual dictionaries

   We do not consider how large the academic fields covered by our bilingual dictionaries are because it is not practical to cover every single correct translation in the source bilingual corpus by hand. In addition, the source bilingual corpus may miss some translations used in practice. We thus evaluate our bilingual dictionaries by measuring how many translations in our dictionaries are correct.

Although we should evaluate our dictionaries constructed from the documents of every society group, it is difficult for some of the groups due to unavailability of up-to-date dictionaries. We evaluate our dictionaries built from the documents of the Institute of Electronics, Information and Communication Engineers (IEICE, consisting of 33311 documents in English and 54854 in Japanese) and the Information Processing Society of Japan (IPSJ, consisting of 12119 documents in English and 15039 in Japanese) using e-Words [Saito and Incept], an online encyclopedia of information and communication terms. While e-Words is an encyclopedia, the caption of a description consists of the translation of the term. Table 1 shows the numbers of the captions with translation in e-Words. Since we do not include a Japanese compound word consisting of numerals or alphabets, we take the captions with no numerals and alphabets as the correct translations.

Table 1: The Captions of e-Words

	Captions	Captions with no numerals and alphabets
In Japanese	1167	956
In English	1355	1123

Table 2 and 3 depict the number of the words found in both our bilingual dictionaries and e-Words, and the number of the words translated correctly, where A is the number of words in a dictionary, B is the number of words found in both the dictionary and e-Words, C is the number of words correctly translated out of B, D is the ratio of words correctly translated out of B, and E is the ratio of the words correctly translated and preferred most by disambiguation out of the number of words with ambiguous translation consisting of correct translation.

While only a small part of the words are evaluated, about 65%-75% of the words in our dictionaries translate correctly. For the ambiguous words, about 55%-70% of the words are disambiguated correctly. As the number of the ambiguous words in our dictionaries are much smaller than the total number of the words, about 80% of the words in our dictionaries are expected to be correct after disambiguation.

Table 2: Precision of our dictionaries from Japanese to English

Source corpus	IEICE	IPSJ	IEICE + IPSJ
A	32376	16533	44202
B	221	208	261
C	142	135	168
D	64.3%	64.9%	64.4%
E	70.4%	75.0%	70.4%

Table 3: Precision of out dictionaries from English to Japanese

Source corpus	IEICE	IPSJ	IEICE + IPSJ
A	22784	10709	30623
B	214	210	247
C	162	160	192
D	75.7%	76.2%	77.7%
E	55.9%	69.4%	56.2%

2. Document retrieval

   We then evaluated our bilingual dictionaries by document retrieval. All of the documents in NTCIR1 test set collection are used to construct bilingual dictionaries. Although our bilingual dictionaries evaluate to have a high precision, the results of document retrieval experiment are rather disappointing. Figure 1 depicts the recall-precision curves of the retrieval results, where 'Japanese-English' means that the queries are in Japanese and the documents are in English, and so on. Queries are generated from the DESCRIPTION field. Partially correct documents are not counted as correct documents. Upon translation of a query, the bilingual dictionary constructed from the documents of the society groups most similar to the query is adopted.

Table 4: The numbers of the words in our dictionaries and indexes

	Dictionaries	Indexes
In Japanese	214296	7111688
In English	188625	2922850

The fatal problem in document retrieval is mismatch of the numbers of the words in our bilin-

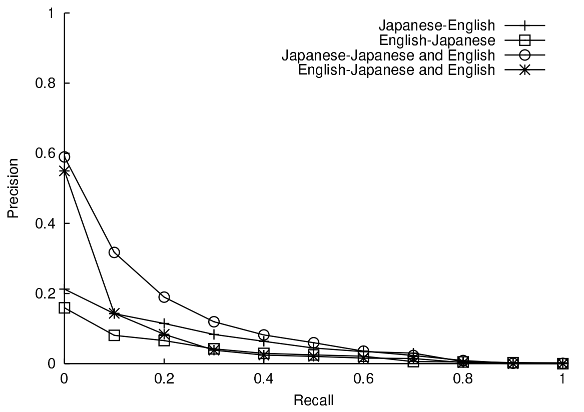

Figure 1: Recall-precision curves of the retrieval results

gual dictionaries and indexes. Table 4 compares the average numbers of the words in our dictionaries per society group, and the numbers of the words in our indexes. Our dictionaries actually cover only 3%-6% of the indexes. Hence most of the index words are not obtained by translation, degrading the retrieval performance hopelessly. The large numbers of the words in our indexes are likely to be caused by extraction of compound words from the documents. We are currently investigating whether we can solve this problem by excluding compound words from our indexes.

7 Conclusion

We described our cross-lingual retrieval system based on construction of bilingual dictionaries from a non-parallel bilingual corpus, using EDICT as the startpoint. We merge compound word translation extracted from a bilingual corpus to the dictionaries. Disambiguation of our bilingual dictionaries is performed using the bilingual corpus. Although the precisions of the constructed bilingual dictionaries are fairly high, the results of document retrieval are poor due to mismatch of the numbers of the words in the dictionaries and indexes.

References

[Breen1995] James W. Breen. 1995. Edict, freeware japanese/english dictionary. ftp://ftp.cc.monash.edu.au/pub/nihongo/00INDEX.html.

[Chen et al. 1999] Aitao Chen, Fredric C. Gey, Kazuaki Kishida, Hailing Jiang, and Qun Liang. 1999. Comparing multiple methods for japanese and japanese-english text retrieval. In *the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pages 49–58, Aug.

[Collier et al. 1998] Nigel Collier, Hideki Hirakawa, and Akira Kumano. 1998. Machine translation vs. dictionary term translation - a comparision for english-japanese news article alignment. In *36th Annual Meeting of the Associatoin for Computational Linguistics and 17th International Conference on Computational Linguistics*, volume 1, pages 263–267.

[Dagan and Church1994] Ido Dagan and Ken Church. 1994. Termight: Identifying and translating technical terminology. In *the 4th Conference on Applied Natural Language Processing*, pages 34–40, Oct.

[Fujii and Ishikawa1999] Atsushi Fujii and Tetsuya Ishikawa. 1999. Cross-language information retrieval at ULIS. In *the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pages 163–169, Aug.

[Fung1995] Pascale Fung. 1995. Compiling bilingual lexicon entries from a non-parallel english-chinese corpus. *Workshop on Very Large Corpora*, pages 173–183.

[Kando et al. 1999] Noriko Kando, Kazuko Kuriya, Toshihiko Nozue, Koji Eguchi, Hiroyuki Kato, and Souichiro Hidaka. 1999. Overview of IR tasks at the first NTCIR workshop. In *the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pages 11–22, Aug.

[Knight and Graehl1998] Kevin Knight and Jonathan Graehl. 1998. Machine transliteration. *Computational Linguistics*, 24(4):599–612, Dec.

[Kupiec1993] Julian Kupiec. 1993. An algorithm for finding noun phrase correspondences in bilingual corpora. In *the 31st Annual Meeting od the Association for Computational Linguistics*, pages 17–22, Jun.

[Lin et al. 1999] Chuan-Jie Lin, Wen-Cheng Lin, Guo-Wei Bian, and Hsin-Hsi Chen. 1999. Description of the NTU japanese-english cross-lingual information retrieval system used for NTCIR workshop. In *the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pages 145–148, Aug.

[Matsumoto et al. 1993] Yuji Matsumoto, Hiroyuki Ishimoto, and Takehito Utsuro. 1993. Structual matching of parallel text. In *the 31th Annual Meeting of the Association for Computational Linguistics*, pages 23–30, June.

[Mitsuhiro et al. 1999] SATO Mitsuhiro, ITO Hayashi, and NOGUCHI Naohiko. 1999. NTCIR experiments at matsushita: Ad-hoc and CLIR task. In *the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pages 71–81, Aug.

[Nakagawa2000] Hiroshi Nakagawa. 2000. Disambiguation of lexical translations based on bilingual comparable corpora. In *Proceedings of Workshop on Terminology Resources and Computation, Second International Conference on Language Resources and Evaluation*, pages 33–38, May.

[Rapp1995] Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. In *the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 320–322, June.

[Saito and Incept] Fumihiko Saito and Incept. Information and communication encyclopedia: e-words. http://www.e-words.ne.jp/ (in Japanese).

[Sakai et al. 1999] Tetsuya Sakai, Yasyuo Shibazaki, Masaru Suzuki, and Masahiro Kajiura. 1999. Cross-language information retrieval for NTCIR at toshiba. In *the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pages 137–144, Aug.

[Smadja et al. 1996] Frank Smadja, Kathleen R. McKeown, and Vasileios Hatzivassiloglou. 1996. Translating collocation for bilingual lexicons: A statistical approach. *Computational Linguistics*, 22(1):1–38, Mar.

[Tanaka and Iwasaki1996] Kumiko Tanaka and Hideya Iwasaki. 1996. Extraction of lexical translation from non-aligned corpora. In *the 16st International Conference on Computational Linguistics*, volume 2, pages 580–585, Aug.

[Tzoukermann et al. 1997] Evelyne Tzoukermann, Judith L. Klavans, and Christian Jacquemin. 1997. Effective use of natural language processing techniques for automatic conflation of multi-word terms: The role of derivational morphology, part of speech tagging, and shallow parsing. In *the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 148–155, Jul.

[Wehrli1998] Eric Wehrli. 1998. Translating idioms. In *the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*, volume 2, pages 1388–1392, August.