Selection of Japanese-English Equivalents by Integrating High-quality Corpora and Huge Amounts of Web Data

Qing Ma[†‡], Nakao Koichi[†], Masaki Murata[‡], Hitoshi Isahara[‡]

[†]Ryukoku University, [‡]National Institute of Information and Communications Technology
[†]Otsu 520-2194, Japan, [‡]Kyoto 619-0289, Japan
qma@math.ryukoku.ac.jp, koichi-n@dsk.zaq.ne.jp, murata@nict.go.jp, isahara@nict.go.jp

Abstract

As a first step to developing systems that enable non-native speakers to output near-perfect English sentences for given mixed English-Japanese sentences, we propose new approaches for selecting English equivalents by using the number of hits for various contexts in large English corpora. As the large English corpora, we not only used the huge amounts of Web data but also the manually compiled large, high-quality English corpora. Using high-quality corpora enables us to accurately select equivalents, and using huge amounts of Web data enables us to resolve the problem of the shortage of hits that normally occurs when using only high-quality corpora. The types and lengths of contexts used to select equivalents are variable and optimally determined according to the number of hits in the corpora, so that performance can be further refined. Computer experiments showed that the precision of our methods was much higher than that of the existing methods for equivalent selection.

1. Introduction

Writing is usually problematic when non-native speakers need to effectively explain ideas or present achievements in English. They usually check the number of hits obtained from Web search engines for English expressions to determine whether these expressions are suitable for the written context. We aimed at developing systems to support English writing that would enable non-native speakers to output near-perfect English sentences for given mixed English-Japanese sentences. As a first step toward developing such systems, we propose several new methods for selecting English equivalents by using the number of hits for various contexts in large English corpora. As the large English corpora, we not only used the huge amounts of Web data but also the manually compiled large, high-quality English corpora with about 9,000,000 English sentences. Using such high-quality corpora enables us to accurately select equivalents, and using huge amounts of Web data enables us to resolve the problem of the shortage of hits that normally occurs when using only high-quality corpora. The contexts, which are also called queries in this paper, consist of the English candidates for the original Japanese words and their neighboring (contextual) English words. The types and lengths of contexts used to select equivalents are variable and optimally determined according to the number of hits in the corpora, so that performance can be further refined.

Computer experiments on 143 test problems that were randomly generated from an English-Japanese corpus, each of which had an average of 15 English equivalent candidates, revealed that an equivalent-selection method using variable context had a precision of 52.45%. The precision was determined by a rigorous criterion of accuracy evaluation. This figure was about 10% (average) and 4% (best case) higher than the precisions of the existing methods. Furthermore, integrating high-quality corpora and huge amounts of Web data achieved a precision of 58.04%, which was the highest out of all the methods proposed in this paper.

2. Related work

The related work can be classified roughly into two categories.

The first is research on equivalent selection by using statistical translation techniques, including example-based approaches, IBM translation models, N-gram models, and machine learning [e.g., (Fung and Yee, 1998), (Ballesteros and Croft, 1998), (Uchimoto et al., 2003), (Yamamoto and Matsumoto, 2001), (Fujii and Ishikawa, 2000)]. Example-based approaches, IBM models, and machine learning, however, require large comparable or parallel corpora, which are high-cost. On the other hand, N-gram models do not always require comparable or parallel corpora, and they are used as baselines in our methods (for details, see Sec. 4.). In fact, our methods do not require any comparable or parallel corpora, and in our experimental results, they demonstrated higher precision than did the baselines.

The second category of related work is research on developing various kinds of English writing support systems, e.g., systems that can present examples of English sentences including keywords given by users (Ichihara et al., 2005), and systems that can detect certain types of English errors related to the mass-count distinction (Nagata et al., 2006). The early studies most related to our work were done by (Oshika et al., 2005) and (Sato et al., 2006). Their methods, however, used only raw Web data (i.e., they simply checked numbers of hits on the Web). The reliability was thus extremely low because of huge numbers of documents written by non-native speakers. In addition, the contextual types (e.g., lexicon or POS) were fixed, and the contextual length (the number of contextual words) was either fixed by the system or set by the users. Therefore, the precision of selecting equivalents was either low or depended largely on the user's English skills.

In our research, we not only use huge amounts of Web data but also have manually compiled large, high-quality English corpora. Using such high-quality corpora enables us to accurately select equivalents, and using huge amounts of Web data enables us to resolve the problem of the shortage

of hits that normally occurs when using only high-quality corpora. In our methods, the types and lengths of contexts used to select equivalents are variable and optimally determined according to the number of hits in the corpora, so that performance can be further refined.

3. System overview

We outline the processes for the system of selecting equivalents we developed in what follows.

a. Input

The input for the system is a mixed English-Japanese sentence, e.g., "The no-nuke *undou* is as active as ever before", where *undou* (movement) is a Japanese word.

b. Dictionary look-up

Japanese words are extracted from the given sentence and their English candidates are obtained by looking them up in a Japanese-English dictionary. All the English candidates for the Japanese word *undou* such as "motion", " exercise", "sport", "campaign", and "movement" are obtained for the sentence given above.

c. Query construction

Queries such as "no-nuke motion is" and "no-nuke movement is", or "POS:noun motion POS:verb" are constructed that are composed of an English candidate and contextual words or POSs for all English candidates. How queries are constructed is the most important issue in this paper and will be described in detail in Sec. 4..

d. Query search

All queries are searched from the corpora and hits are obtained. To search the Web data, we merely use Google search engine. To search the high-quality corpora, on the other hand, we developed an original system to enable flexible searches in the sense of following points.

1. Wildcard searches

As the case may be, a part of a query in searches should be arbitrary words (strings) to alleviate the problem with the shortage of hits. We adopted wildcard matches to achieve this. For example, the query "read *2 book" can match any pattern bounded by "read " and "book" and they are separated by up to two words.

2. POS searches

As the case may be, a part of a query in searches might be POSs. For example, we usually need to eliminate the differences between the articles "a" and "the" so that the problem with the shortage of hits can be reduced. We therefore introduced the POS search function. For example, a query might be "read /DT book" where "/DT" means the POS that is an article.

3. Variation searches

English verbs and nouns have numerous variations according to their person and tense, and whether they are singular or plural. To limit these variations as much as possible, we adopted variation searches. Therefore, the query "read*VB /DT book*NN", for example, can match "He reads the book", "He read a book", and "He read(s) the books".[1]

e. Output

The candidate with which the query is constructed that has the largest number of hits is selected as the equivalent answer and is output based on the number of hits for all queries. "Movement" will be selected in this example.

[1] Variations in irregular verbs cannot presently be handled.

4. Methods of constructing queries

This section first describes existing methods, which have various improvements, as baselines. The baseline methods are not only used for comparisons with ours, but also used in the integrated system, i.e., one of our methods in which both high-quality corpora and huge amounts of Web data are used. We then describe the integrated system, as well as methods where the types and lengths of contexts used for selecting equivalents are not fixed, as they are with the existing methods, but variable and optimally determined according to the state of hits on corpora.

4.1. Baselines

4.1.1. Use of candidate words only

Queries are constructed using only the candidate word, i.e., the candidate with the largest number of hits on the corpora is selected as the equivalent answer.

4.1.2. Use of fixed-length word sequences (i.e., N-gram models)

Queries are constructed using a fixed-length word sequence centered on the candidate word, with (l, r) contextual words to the left and right, respectively. If the fixed-length word sequence is too long, however, the number of hits on the corpora will significantly decrease and the results for selecting equivalents will worsen. We introduced two rules to reduce this.

Rule 1

If a contextual word is a punctuation mark such as "." or "," and it is to the left (right) of the candidate, then the mark and the words to its left (right) will not be included in the query.

Rule 2

Contextual words two words away from the candidate are replaced by their POSs when using high-quality corpora for searches.

For example, for sentence "He didn't know the outcome of the meeting" where "outcome" is a candidate, the query constructed is "/VB the outcome of /DT" when $(l, r) = (2, 2)$.

This method is equivalent to a type of N-gram smoothing model: in the case of $(l, r) = (1, 0)$, it is equivalent to the bi-gram model; and in the case of $(l, r) = (2, 0)$, it is equivalent to the tri-gram model. In fact, our experimental results showed that only very few queries constructed with this method for the bi-gram and tri-gram models could not obtain hits in the high-quality corpora, which means that the data sparseness problem has been resolved in these cases (see the no-hits ratio of (l, r)=(1,0) and (l, r)=(2,0) in Table 2).

[2] The intention in replacing words with a "high frequency of appearance" not words with "low frequency of appearance" is that we want to replace the verb "be" and its variations, the articles "the" and "a", and the pronouns such as "he" and "she" with their POSs, which are usually those with a high frequency of appearance.

Step 0
Set l and r to large values, where l and r are the number of contextual words to the left and right of the candidate.
Step 1
Construct a query with the candidate word and words to its l left and r right and search the corpora.
Step 2
When there is a hitting, select the candidate word with which the query has the largest number of hits as the answer to the selection of equivalents and terminate processing. Otherwise, go to Step 3.
Step 3
The $(l + r)/2$ words in the query are replaced by their POSs and the corpora are searched, where words with a high frequency of appearance are selected to be replaced[2].
Step 4
Once a hit occurs, select the candidate word with which the query has the largest number of hits as the answer to the selection of equivalents and terminate processing. Otherwise, go to Step 5.
Step 5
If there are fewer hits to the left of the query than those to the right, the l is decremented by one. Otherwise, the r is decremented by one. Here, the part to the left (right) is constructed with a sequence bounded by the left-most (right-most) word and the candidate; the words wedged between them are replaced by their POSs.
Step 6
If $l > 0$ or $r > 0$ go to Step 1, otherwise terminate processing.

Table 1: Algorithm for constructing variable-length queries

4.1.3. Use of fixed-length POS sequences

This follows the same procedures as that in Sec. 4.1.2. but replaces the words with their POSs; however, it can only be used when using high-quality corpora for searches.

4.1.4. Use of contentive-bounded word sequences

In the method proposed by (Sumita et al., 2004), all queries are constructed using the shortest word sequence centered on the candidate word and containing contextual words wedged between contentives (i.e., nouns, verbs, and adjectives). Since this method usually tends to construct long queries leading to a shortage of hits, we replaced non-contentive words in the context with their POSs when using high-quality corpora for searches. For example, for the sentence "He didn't know the outcome of the meeting" where "outcome" is a candidate, the query is "know /DT outcome /IN /DT meeting", which is wedged between the contentives, "know" and "meeting" and includes POSs, /DT and /IN.

4.2. Proposed methods

4.2.1. Rule-based method

When non-native speakers use the number of hits of English expressions on the Web to determine their suitability of their written English, several tendencies in constructing queries can be observed. For example, when the candidate is a noun, a set consisting of a verb and a dependent noun is typically used. In this method, we construct queries by using rules to representing these tendencies, according to POS tagging and dependency analysis.

Rules for verb candidates

1. Check the words to the right of the candidate from left to right. Once a noun is found, a query is then constructed with a word sequence bounded by the candidate and the noun. The words wedged between them are replaced by wildcard asterisks.

For example, for the sentence "The enterprise will go bankrupt soon" , the query is "go bankrupt" and for "The enterprise will go into bankrupt", the query is "go * bankrupt", where "go" is a candidate.

2. If no nouns are found, then check words to the left of the candidate from right to left. Once a noun is found, a query is then constructed with the word sequence bounded by the noun and the candidate. The words wedged between them are replaced with wildcard asterisks. If there are no nouns either to the right or the left, then only the candidate is used to construct the query.

For example, for the sentence "The rate is reversed with 41% thinking "there will be a negative impact" and 52% "there will not.", where "reversed" is a candidate, then query is "rate * reverse*VB" (for the meaning of *VB see **d. Query search** of Sec. 3.).

Rules for non-verb candidates

1. If the phrase including the candidate is a noun phrase (NP), then all the words in the NP are used to construct the query. However, if all the words other than the candidate in the NP are articles, e.g., "the book" (we call this case an "exception" here), then the next rule is adopted.

For example, for the sentence "He knew the outcome of the meeting" where "outcome" is a candidate, then the query is "the outcome of the meeting".

2. If the phrase including the candidate is not an NP, or is the exception mentioned above, then the positions of the verbs to the left and right of the candidate are checked and the query is constructed with a word sequence bounded by the candidate and the closest verb. The words wedged be-

tween them are replaced by wildcard asterisks.
For example, for the sentence "He read the book which his father gave him" where "book" is a candidate, then the closest verb "read" is selected and the query is "read * book".

4.2.2. Variable-length context method

Generally, the longer the query, the better its reliability, and long queries should be used as much as possible. As the length of a query is increased, however, the number of hits in the corpora significantly decreases, and the results for selecting equivalents are degraded. In this method, the query is initially long and consists of words only. The words are replaced by their POSs or deleted in stages, starting far from the candidate and moving closer as needed, so that the length of the query is gradually reduced. The algorithm for constructing variable-length queries is shown in Table 1. However, it can only be used when using high-quality corpora for searches.

4.2.3. Integrated system

The high-quality English corpora are first used to select equivalents for the Japanese words. The huge amounts of Web data is then used to alleviate shortages of hits. The details will be described in the next section.

5. Experimental results

5.1. Data

We collected high-quality corpora made up of a total of 8,938,326 sentences (176,272,651 words) that we used in the experiments in advance. They consisted of the Wikipedia data including approximately 2,000,000 sentences downloaded from "Wikimedia Downloads" site[3], the British National Corpus (BNC) including 6,050,000 sentences, and others including the three years (2003-2005) of "The Daily Yomiuri" English newspaper and the automatically aligned Japanese-English sentences (Utiyama and Isahara, 2003). On the other hand, the Web data were not gathered beforehand but used to check hits on the Web using Google Web APIs[4].
Eijiro[5] with 1,760,000 lexical entries was used as the Japanese-English dictionary. The SS Tagger (Tsuruoka and Tsujii, 2005a) was used for tagging English POSs. The SS Parser (Tsuruoka and Tsujii, 2005b) was used for parsing English. Chasen (Matsumoto et al., 2000) was used for segmenting Japanese words and tagging POSs.
The testing problems were generated as follows. One-hundred-fifty English sentences were randomly selected from the NICT Japanese-English parallel corpus in which the original Japanese sentences were Japaneses newspaper articles and the English sentences were translated by professional translators (Uchimoto et al., 2004). One word was randomly selected for each of these sentences and replaced with its Japanese equivalent appearing in the original Japanese sentence, so that in total there were 60 Japanese words whose English equivalents were nouns, 60 whose equivalents were verbs, and 30 whose equivalents were adjectives. It should be noted that we ended up with only 143 testing problems for the experiments, because the remaining seven problems were "no-solution problems"[6]. Each Japanese word in these testing problems had 15 equivalent candidates on average.

5.2. Accuracy Evaluation

Ideally, language experts should have judged whether answers were correct. Judging dozens of possible outputs on 143 problems, however, would be an extremely time-consuming endeavor. The fixed-length word method, for example, has a total 15 combinations of (l, r) when varying both parameters from 0 to 3. At the present stage, we therefore only considered English equivalents as correct if they were exactly the same as the words in the Japanese-English parallel sentences of the NICT corpus.
Considering that the problem with the shortage of hits might be resolved by using huge amounts of Web data, we calculated two kinds of precisions:

$$P(\text{with no-hits}) = \frac{\#(\text{correct answers})}{\Sigma} \times 100\%$$

and

$$P(\text{w/o no-hits}) = \frac{\#(\text{correct answers})}{\Sigma - \#(\text{no-hit problems})} \times 100\%$$

where Σ is the total number of testing problems and $\#$ means number. In this paper, we only considered cases where no patterns on the searched corpora could be matched by queries as shortage of hits and these are called "no-hits" later. To see how each method finds a hit in the corpora, we calculated the "no-hits ratio" as follows.

$$\text{no-hits ratio} = \frac{\#(\text{no-hit problems})}{\Sigma} \times 100\%$$

5.3. Results

For the fixed-length methods, both parameters of (l, r) were varied from 0 to 3. Table 2 and Table 3 list the precisions of the total combination of (l, r) for the fixed-length word methods and the fixed-length POS methods with the high quality corpora. Table 4 lists the precisions of the total combination of (l, r) for the fixed-length word method with Web data. From these tables we can see that the highest precisions of the fixed-length methods were 48.95%, 48.25, and 49.65%, and the average precisions of the fixed-length methods were 39.30%, 41.82%, and 36.78%, respectively. Furthermore, an optimal pair of (l, r) could not be found that had high precisions when both the high quality corpora and Web data were used.
Table 5 compares the precisions obtained using the various methods with the high-quality corpora, in which the average values and the best figures for the fixed-length methods from the total 15 combinations of (l, r) are shown. For the variable-lengths method, the initial values of (l, r) were set to (3, 3). The results show that the proposed variable-lengths method and the rule-based method had the highest and the second highest precisions, respectively.

[3] http://download.wikimedia.org/
[4] http://code.google.com/apis.html
[5] http://www.eijiro.jp/index.html
[6] "No-solution problems" are where Japanese words obtained by Chasen analyses cannot presently be correctly looked-up in the Japanese-English dictionary.

(l, r)	P(with no-hits)	P(w/o no-hits)	no-hits ratio
(0, 1)	45.45	46.76	2.67
(0, 2)	44.06	47.01	6.00
(0, 3)	43.36	49.21	11.33
(1, 0)	48.25	50.00	3.33
(1, 1)	43.36	52.99	17.33
(1, 2)	39.16	53.85	26.00
(1, 3)	37.06	55.21	31.33
(2, 0)	**48.95**	51.47	4.67
(2, 1)	44.06	58.33	23.33
(2, 2)	37.06	57.61	34.00
(2, 3)	31.47	54.88	40.67
(3, 0)	43.36	49.60	12.00
(3, 1)	32.17	52.87	37.33
(3, 2)	27.27	52.00	45.33
(3, 3)	24.48	52.24	50.67
avg.	39.30	52.27	23.07

Table 2: Precision and no-hit ratio (%) for fixed-length words method with high-quality corpora

(l, r)	P(with no-hits)	P(w/o no-hits)	no-hits ratio
(0, 1)	45.45	45.77	0.67
(0, 2)	45.45	46.10	1.33
(0, 3)	45.45	47.10	3.33
(1, 0)	41.96	41.96	0.00
(1, 1)	46.85	47.52	1.33
(1, 2)	44.76	47.06	4.67
(1, 3)	45.45	51.59	11.33
(2, 0)	44.76	45.39	1.33
(2, 1)	**48.25**	50.74	4.67
(2, 2)	40.56	46.77	12.67
(2, 3)	33.57	45.71	25.33
(3, 0)	42.66	45.19	5.33
(3, 1)	41.26	47.97	13.33
(3, 2)	34.27	47.57	26.67
(3, 3)	26.57	48.10	42.67
avg.	41.82	46.97	10.31

Table 3: Precision and no-hit ratio (%) for fixed-length POS method with high-quality corpora

Table 6 lists the precisions for the various methods with Web data. In this case, the fixed-length POSs and variable-lengths methods are not applicable, because we merely checked hits on the Web with Google searches and there was no way of matching POS information between the queries and the Web data. From the table, we see that the best figure from the total 15 combinations of (l, r) of the fixed-length word method had the highest precision. However, this figure was lower than that obtained when the proposed variable length method was used to search the high quality corpora.

Table 7 lists the precision for the integrated system. In this system, the contentive bounded method, which had the highest precision when using high-quality corpora and not counting the no-hit problems (i.e., P(w/o no-hits) of Table 5), was first adopted on the high-quality corpora. This obtained 32 correct answers and generated 91 cases with no-hits. For the 91 cases with no-hits, the fixed-length words method of $(l, r) = (0.3)$, which had the highest precision when using Web data (i.e., P(with no-hits) of Table 6), was therefore adopted on Web data. As a result, we finally obtained 83 correct answers and a precision of 58.04%, which was the highest of the tested methods[7].

(l, r)	P(with no-hits)	P(w/o no-hits)	no-hits ratio
(0, 1)	44.06	44.06	0.00
(0, 2)	49.65	49.65	0.00
(0, 3)	**49.65**	52.99	6.00
(1, 0)	25.87	26.06	0.67
(1, 1)	43.36	44.93	3.33
(1, 2)	46.15	50.77	8.67
(1, 3)	39.16	52.34	24.00
(2, 0)	31.47	32.14	2.00
(2, 1)	39.86	46.34	13.33
(2, 2)	37.76	56.25	31.33
(2, 3)	27.97	61.54	52.00
(3, 0)	36.36	42.28	13.33
(3, 1)	37.06	54.08	30.00
(3, 2)	25.17	56.25	52.67
(3, 3)	18.18	70.27	70.67
avg.	36.78	49.33	20.53

Table 4: Precision and no-hit ratio (%) for fixed-length words method with Web data

Methods	P(with no-hits)	P(w/o no-hits)	no-hits ratio
Candidates only	40.56	40.56	0.00
Fixed-length words (avg.)	28.07	52.93	44.67
Fixed-length words (2,0)	48.95	51.47	4.67
Fixed-length POSs (avg.)	41.82	46.97	10.31
Fixed-length POSs (2,1)	48.25	50.74	4.67
Contentive bounded	22.38	**61.54**	**60.67**
Rule-based	**49.65**	55.04	9.33
Variable lengths	**52.45**	52.45	0.00

Table 5: Precision (%) for various methods with high-quality corpora

6. Conclusion

As a first step in developing systems to support English writing that enable non-native speakers to output near-perfect English sentences for given mixed English-Japanese sentences, we have developed several new methods for selecting English equivalents to Japanese words by using the number of hits for various contexts in large En-

[7] We realize that the experimental results for the integrated system were somewhat closed because we used the best methods for the two kinds of corpora, which were preliminarily known throughout the experiments. Open tests should be done by using new testing problems in the future.

Methods	P(with no-hits)	P(w/o no-hits)	no-hits ratio
Candidates only	40.56	40.56	0.00
Fixed-length words (avg.)	36.78	49.33	20.53
Fixed-length words (0,3)	**49.65**	52.99	6.00
Fixed-length POSs	-	-	-
Contentive bounded	39.16	54.37	26.67
Rule-based	37.06	37.86	2.00
Variable lengths	-	-	-

Table 6: Precision (%) for various methods with Web data

Method	P(with no-hits)
Integrated system (Contentive bounded + Fixed-length words (0,3))	**58.04**

Table 7: Precision (%) for integrated system using both high-quality corpora and Web data

glish corpora. As the large English corpora, we not only used the huge amounts of Web data but also the manually compiled large, high-quality English corpora. Using the high-quality corpora enabled us to accurately select equivalents, and using huge amounts of Web data enabled us to resolve the problem of the shortage of hits that normally occurs when using only high-quality corpora. Computer experiments on 143 test problems demonstrated that the proposed variable-length context method had a precision of 52.45%, which was about 10% (average) and 4% (best case) higher than the precisions of existing methods. The rule-based method had the second highest precisions when using the high-quality corpora. The variable-lengths method with the high-quality corpora had the highest precision when using either the high-quality corpora or the Web data. An integrated system using these methods with both the high-quality corpora and Web data achieved a precision of 58.04%, which was the highest out of all the methods detailed in this paper. Given that each test problem had an average of 15 English equivalents and that the criterion for judging correct answers was extremely rigorous, these results are encouraging.

We intend to perform further experiments on larger-scale testing problems to confirm how effective the proposed methods are and to further improve them in future work. The objects to support English writing will also be expanded to the level of the expressions including phrases, clauses, and sub-sentences from the word level.

7. Acknowledgements

This work was supported by Grant-in-Aid Scientific Research (KAKENHI (C) 19500133), The Ministry of Education, Culture, Sports, Science and Technology, Japan.

8. References

L. Ballesteros and W. B. Croft. 1998. Resolving ambiguity for cross-language retrieval. In *Proceedings of 21th ACM SIGIR*, pages 64–71.

K. Fujii and T. Ishikawa. 2000. Translating compound words in the cross-language information retrieval of technical documents (in Japanese). In *Journal of Information Processing Society of Japan*, pages 1038–1045.

P. Fung and L. Y. Yee. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of COLING-ACL98*, pages 414–419.

K. Ichihara, S. Tada, S. Mizobuti, and K. Ando. 2005. English verification and example retrieval tool for English writing(in Japanese). In *The 11th Annual Meeting of the Association for Natural Language Processing (NLP2005)*, pages 4–9.

Y. Matsumoto, A. Kitauchi, T. Yamashita, Y. Hirano, H. Matsuda, K. Takaoka, and M. Asahara. 2000. Japanese morphological analysis system chasen version 2.2.1.

R. Nagata, T. Wakana, A. Kawai, K. Morihiro, F. Masui, and N. Isu. 2006. Recognition errors in English writing based on the mass count distinction (in Japanese). In *IEICE Transaction on Information and Systems, Vol. J89-D, No. 8*, pages 1777–1790.

H. Oshika, M. Sato, S. Ando, and H. Yamana. 2005. An English composition support system using Google (in Japanese). In *DEWS2005 4-B-08*.

M. Sato, S. Ando, and H. Yamana. 2006. Constuction of English composition support systems using search engin (in Japanese). In *The 12th Annual Meeting of the Association for Natural Language Processing (NLP2006)*, pages 664–667.

E. Sumita, F. Sugaya, and S. Yamamoto. 2004. Automatic generation method of a fill-in-the-blank question for measuring English proficiency (in Japanese). In *ICICE TL2004-22/WIT2004-56*, pages 17–22.

Y. Tsuruoka and J. Tsujii. 2005a. Bidirectional inference with the easiest-first strategy for tagging sequence data. In *HLT/EMNLP 2005*, pages 467–474.

Y. Tsuruoka and J. Tsujii. 2005b. Chunk parsing revisited. In *the 9th International Workshop on Parsing Technologies (IWPT 2005)*, pages 133–140.

K. Uchimoto, S. Sekine, M. Murata, and H. Isahara. 2003. Word translation by combining and example-based method and machine learning models (in Japanses). In *Natural Language Processing, Vol. 10, No.3*, pages 87–114.

K. Uchimoto, Y. Zhang, K. Sudo, M. Murata, S. Sekine, and H. Isahara. 2004. Multilingual Aligned Parallel Treebank Corpus Reflecting Contextual Information and Its Applications. In *Proceedings of the MLR2004: PostCOLING Workshop on Multilingual Linguistic Resources*, pages 63–70.

M. Utiyama and H. Isahara. 2003. Reliable measures for aligning Japanese-English news articles and sentences. In *ACL-2003*, pages 72–79.

K. Yamamoto and Y. Matsumoto. 2001. Translation pattern acquisition using dependency structures (in Japanese). In *Journal of Information Processing Society of Japan*, pages 2239–2247.