SENSEVAL-2 Japanese Translation Task

Sadao Kurohashi† and Kiyotaka Uchimoto††

This paper describes the SENSEVAL-2 Japanese translation task. In this task, word senses are defined according to distinct translations in a given target language. A translation memory (TM) was constructed which contains, for each Japanese head word, a list of typical Japanese expressions and their English translations. For each test word instance, participants were required to submit the TM record best approximating that usage, or alternatively, actual target word translations. There were 9 system entries from a total of 7 organizations.

KeyWords: SENSEVAL, *Word Sense Disambiguation, Translation Memory, Machine Translation*

1 Introduction

In written texts, words which have multiple senses can be classified into two categories: homonyms and polysemous words. Generally speaking, while distinctions between homonyms are quite clear, polysemous senses are very subtle and hard to distinguish (Kilgarriff and Palmer 2000). English texts contain many homonyms. In Japanese texts, on the other hand, most content words are written as ideograms and there are rarely homonyms.[1] That is, the main target in Japanese WSD is polysemy, which makes the Japanese WSD task very hard to set up. The sense distinction of polysemous words depends heavily on how those senses are to be used, that is, the application of WSD (Ide 2000).

Given this setting, for SENSEVAL-2 (the second in a series of evaluation exercises for WSD programs) a Japanese translation task was organized, in addition to a conventional Japanese dictionary task. In the translation task, word sense is defined according to translation distinction. Here, we set up the task based on the example-based machine translation paradigm (Nagao 1981; Pinkham, Och, and Knight 2001). That is, first, a translation memory (TM) is constructed which contains, for each Japanese head word, a list of typical Japanese expressions (phrases/sentences) involving the head word and an English translation for each (Figure 1).

† University of Tokyo
†† Communications Research Laboratory
The task was designed by both authors of this paper. The first author then acted as the organizer of the task, and the second author became one of the task participants.

1 In case of the spoken language or texts written in kana (the Japanese syllabary), Japanese has a lot of troublesome homonyms, e.g. *kouen* which can mean 講演/lecture, 公園/park, 後援/support, 公演/public performance, and so on.

無理 *muri*	
参加は無理だ	It is impossible to participate.
今から図書館の利用は無理だ	It is impossible to make use of the library in this hour.
今回の法案には無理がある	This bill is hard to pass.
彼が怒るのも無理はない	It is no wonder he got angry.
一番無理のない方法	the most natural way
無理を重ねる	to work too much
無理な話	unreasonable demand
無理な追い越し	passing by force
無理心中を図る	to commit a forced double suicide
...	...

Fig. 1 A section of the translation memory.

We term each pairing of a Japanese and English expression in the TM a TM record. Given an evaluation document containing a target word, participants have then to submit the TM record best approximating that usage.

Alternatively, submissions can take the form of actual target word translations, or translations of phrases or sentences including each target word. This allows existing rule-based machine translation (MT) systems to participate in the task, and allows us to compare TM-based systems with existing MT systems.

For evaluation, we used newspaper articles. The number of target words was 40, and 30 instances of each target word were provided, making for a total of 1,200 instances.

The time schedule for the SENSEVAL-2 Japanese translation task was as follows:

2000/2/23	Call for expressions of interest
2001/1/31	Trial data available
2001/3/16	Translation memory available
2001/5/11	Test data available
2001/6/1	Deadline to submit answers
2001/7/6,7	Workshop (preceding ACL-2001) and notification of results

2 Construction of the Translation Memory

The translation memory (TM) was constructed in two steps:

(1) By referring to the KWIC (Key Word In Context) concordance lines for each target word, typical Japanese expressions were manually extracted by lexicographers.

(2) The Japanese expressions were translated by a translation company.

Phrase uni-gram	Phrase bi-gram		Phrase tri-gram
597 無理な	151 無理は ない。	19 ことには 無理が	7 ことには 無理が ある。
551 無理が	138 無理が ある。	14 とても 無理。	6 求めるのは 無理が ある。
416 無理やり	106 無理も ない。	13 ことは 無理と	5 ことには 無理からぬ 理由が
413 無理に	101 無理 なく	10 求めるのは 無理が	5 嘆くのも 無理は ない。
403 無理を	67 無理の ない	10 とても 無理」と	5 同署は 無理心中と みている。
351 無理。	56 無理が ある」と	9 いうのは 無理が	4 しても 無理は ない。
...	...	...	...

Fig. 2 Example KWIC concordance lines (numbers indicate phrase frequency).

KWIC concordance lines were constructed from the nine years' worth of articles in the Mainichi Newspaper corpus. These were morphologically analyzed and segmented into phrase sequences, and then the 100 most frequent phrase unigrams, bigrams (two types, with the target word in the first or second phrase) and trigrams (the target word is in the middle phrase) were provided to lexicographers (Figure 2).

The lexicographers extracted typical expressions associated with each target word from the KWIC concordance lines. If the sense of the target word was clear from the limited word context, the expression was adopted as is. If its sense was not clear, some pre/post expressions were supplemented by referring to the original newspaper corpus.

Next, we asked a translation company to translate the Japanese expressions. As a result, a TM containing 320 head words and 6920 records was constructed (one head word has 21.6 TM records on average). The average number of words contained in each Japanese expression is 4.5.

3 Gold Standard Test Data and Evaluation of the Translations

As gold standard test data for the task, 40 target words were chosen out of the 320 target words contained in the TM. In order to achieve basic comparability with the dictionary task, the 40 target words were chosen to be a subset of the 100 target words in the dictionary task.

In the Japanese dictionary task, target words were classified into three classes based on the entropy of word sense distribution $E(w)$ in the training data (Shirai 2003). Obviously, the higher $E(w)$ is, the more difficult it becomes to disambiguate w. The three classes were defined as follows:

$C_{difficult}$: $E(w) \geq 1$,

$C_{intermediate}$: $0.5 \leq E(w) < 1$, and

Table 1 The 40 target words for evaluation.

	Noun	Verb
$C_{difficult}$	一般/general(33) 近く/nearby(15) 姿/look(28) 意味/meaning(22) 胸/chest(30)	出る/go out(30) 受ける/receive(22) 持つ/have(59) 聞く/hear(25) 与える/give(36)
$C_{intermediate}$	前/foward(25) 時代/era(39) 今/now(15) 場合/case(23) 事業/business(17) 国内/domestic(14) 一方/one hand(14) 言葉/word(35) 記録/record(18) 市民/citizen(23)	言う/say(32) 使う/use(19) 作る/make(19) 書く/write(15) 超える/exceed(14) 伝える/tell(19) 守る/protect(16) 待つ/wait(23) 描く/draw(12) 乗る/ride(23)
C_{easy}	問題/problem(32) 核/nuclear(8) 中心/center(15) 反対/opposition(26) 花/flower(27)	求める/request(10) 認める/admit(10) 見せる/show(20) 買う/buy(31) 図る/promote(17)

English glosses indicate one possible translation. The figures in parentheses show the number of TM records associated with each target word.

C_{easy}: $E(w) < 0.5$.

The 40 target words in the translation task are made up of 20 nouns and 20 verbs: 5 nouns and verbs from $C_{difficult}$, 10 nouns and verbs from $C_{intermediate}$, and 5 nouns and verbs from C_{easy} (Table 1).

For each target word, 30 instances were chosen from the Mainichi Newspaper corpus (making up a total of 1,200 instances), which again overlapped with the dictionary task. Since the dictionary task uses 100 instances for each target word, the translation task used the 1st, 4th, 7th, ... 90th instances from the dictionary task. Each instance was given in the context of the full newspaper article containing that instance.

As gold standard test data, zero or more appropriate TM records were assigned to each instance by the same translation company as translated the TM records. Translation-equivalent TM records were classified according to the following three classes:

◎ : The translation associated with the TM record is appropriate for the instance. Note that there may be a mismatch in POS, tense, number (singular/plural), or subtleties in the nuance of the translation.

○ : If the instance is considered alone, the English translation associated with the TM record is correct, but in the given context, the translation leads to an awkward or circuitous translation. For example, suppose the target instance is 使用した in 砂糖を使用した飲み物, whose natural translation is "a drink with sugar." The TM record, ナイフを使用する/to use a knife, can be used to translate the instance, but it causes the circuitous translation "a drink which uses sugar."

△ : When considered alone, the English translation is correct, but in the given context,

the translation associated with the TM record is inappropriate.

Out of 1,200 instances, 34 instances (2.8 %) were assigned no TM records (i.e. no TM record provided an appropriate translation at any level). To one instance, an average of 6.6 records were assigned as ◎, 1.4 records as ○, and 0.1 records as △, resulting in a total of 8.1 records. If a system were to, for each instance, randomly output a single TM record associated with the given target word, the accuracy would be 36.8 % in the case that all ◎, ○ and △ TM records were regarded as correct, and 29.0 % in the case that only ◎ records were regarded as correct. These comprise the baseline scores described in the next section.

A single translator annotated all of the 1,200 instances, and a second translator annotated 90 of the instances (9 words × 10 instances), so as to check the level of annotator agreement. For each instance, one record was chosen randomly from annotator B's set of acceptable TM records, and checked against annotator A's TM record annotation. Agreement was 86.6 % in the case that all of ◎, ○ and △ records are regarded as correct, and 80.9 % in the case that only ◎ is regarded as correct.

For submissions in the form of translation data, translation experts (from the same company as constructed the TM and the gold standard test data) were asked to rank the supplied translation as ◎, ○ or incorrect (×). This evaluation does not reflect the overall translation quality of a given sentence/article, but just the appropriateness of the target instance translation.

4 Results

4.1 Participant systems

A total of 9 systems from 7 organizations participated in the translation task. The characteristics of the various systems are summarized as follows:

- AnonymX, AnonymY
  Commercial, rule-based MT systems.
- CRL-NYU (Communications Research Laboratory & New York Univ.) (Uchimoto, Sekine, Murata, and Isahara 2003)
  TM records are classified according to the English head word, and each cluster is supplemented by several corpora. The system returns a TM record when the similarity between a TM record and an input sentence is very high. Otherwise, it returns the English

head word of the most similar cluster by using several machine learning techniques.[2]

- Ibaraki (Ibaraki Univ.) (Shinnou 2001)
  Training data was manually constructed from newspaper articles, with 170 instances for each target word. Features were collected in a 7-word window around the target word, and a decision list was used for learning.
- Stanford-Titech1 (Stanford Univ. & Tokyo Institute of Technology) (Baldwin, Okazaki, Tokunaga, and Tanaka 2001)
  The system selects the appropriate TM record based on Dice's coefficient applied over character-bigrams in a 21-character window around each target word. Inter-instance similarity is also used.
- AnonymZ
  Each sentence (TM records for learning, and instances for testing) is morphologically analyzed and converted into a semantic tag sequence; maximum entropy was used for learning.
- ATR (Kumano, Kashioka, and Tanaka 2003)
  The system selects the most similar TM record based on the cosine similarity between context vectors, which were constructed from semantic features and syntactic relations of neighboring words of the target word.
- Kyoto (Kyoto Univ.)
  The system selects the most similar TM record by a flexible matching algorithm which copes with several types of lexical and syntactic paraphrases.
- Stanford-Titech2 (Stanford Univ. & Tokyo Institute of Technology) (Baldwin et al. 2001)
  The system selects the appropriate TM record based on the case-frame-based similarity, using NTT Goi-Taikei thesaurus.

4.2 Evaluation Results

The results for the respective systems are shown in Figure 3. The left bars indicate system accuracy based on lenient evaluation criteria (all of ◎, ○ and △ are regarded as correct in TM record selection, and ◎ and ○ are regarded as correct for the MT systems); the right bars indicate system accuracy based on strict evaluation criteria (◎ only regarded as correct

2 When the CRL-NYU system returned English head words, it was evaluated by the translation experts as for the MT systems. However, since the basic architecture is to select a translation pattern from the extended TM, we classify it as a TM selection system in the following discussion.

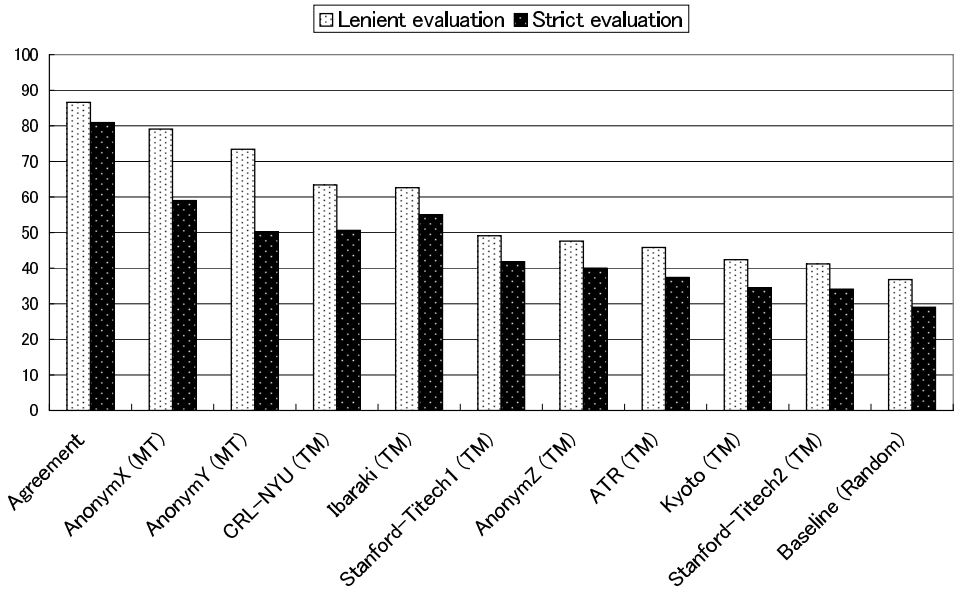

Fig. 3 Results for the Japanese translation task.

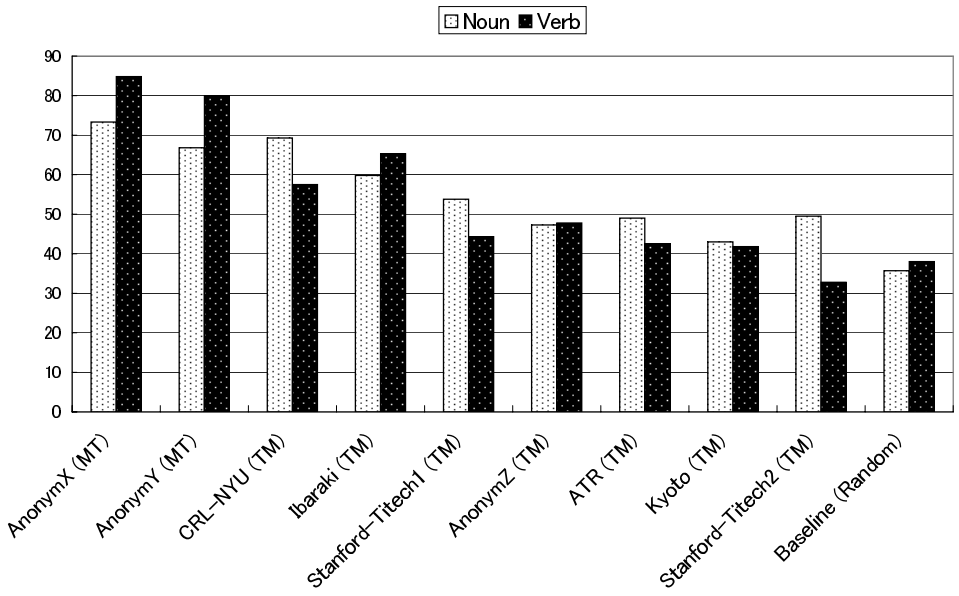

Fig. 4 Scores for nouns and verbs.

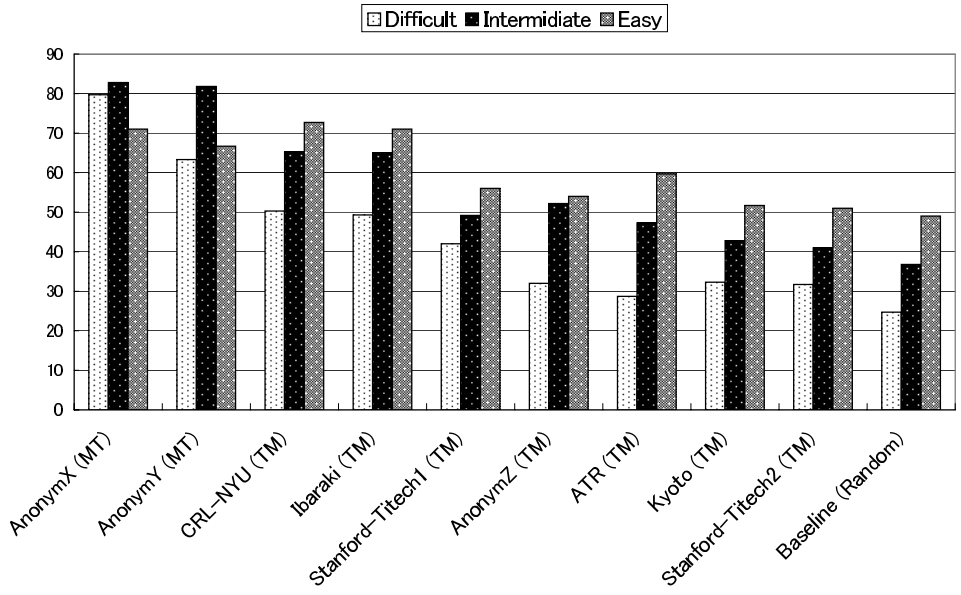

Fig. 5 Scores for the different difficulty classes.

for both TM record selection and the MT systems). [3] [4]

"Agreement" and "Baseline" in Figure 3 are as described in the previous section. If a system determines that there is no appropriate TM record for a given instance, it can return "UNASSIGNABLE". In this case, if there is no appropriate TM record assigned in the gold standard test data, the answer is regarded as correct.

As shown in Figure 3, among the TM selection systems, the systems using some extra learning data, namely CRL-NYU and Ibaraki, outperformed the other systems just using the TM data.

Figure 4 shows scores for nouns and verbs separately, evaluated according to the lenient criteria. While both MT systems could handle verbs better than nouns significantly, the TM record selection systems were variable in relative performance over the two parts of speech. We return to discuss this issue below.

Figure 5 shows scores (based on the lenient criteria) for difficult/intermediate/easy cases.

3 Since the TM does not have a hierarchical structure, it is not possible to utilize evaluation options such as fine, coarse, and mixed, as are used for other SENSEVAL-2 tasks.

4 Note that some participants expected that a select few best-matching TM records would be chosen for each instance in the gold standard test data, which they would have to pinpoint accurately. In fact, however, all possibly useful TM records were marked as correct in the gold standard test data (on average, 8.1 records per instance). Some participants claimed that if they had known that so many records were to be marked correct, alternative learning technique could have been feasible.

Table 2 Comparison of the two MT systems and the two best TM record selection systems ("○" = system output correct, "×" = system output incorrect).

	TM:○○	TM:○×	TM:××	(total)
MT:○○	376	335	98	809
MT:○×	90	75	47	212
MT:××	55	60	64	179
(total)	521	470	209	1200

All the TM record selection systems performed approximately equivalently over the difficulty cases. This suggests that difficulty in machine translation (defined in terms of TM record selection) is heavily correlated with difficulty in monolingual WSD, which is in some sense reasonable. On the other hand, there is no correlation between the performance of the MT systems over the relative levels of difficulty. This is probably because the MT systems have been incrementally improved over time, and fine-tuned to cope with the more difficult expressions.

4.3 Comparison of MT systems and TM record selection systems

Comparison between the MT systems and the TM record selection systems is not easy. Since the method used to evaluate the respective system types was different, their scores cannot be compared in a simplistic way. Furthermore, as shown in Figure 4 and 5, the relative performance of the two system types was quite different.

In order to get a better grasp on the relative results, we classified the 1200 test instances according to the correctness of output of the two MT systems and the two best TM record selection systems (CRL-NYU and Ibaraki). The results are shown in Table 2. From among these, we hand-checked the 55 instances which were correctly handled by both TM systems but neither of the MT systems, and the 98 instances which were correctly handled by both MT systems but neither TM record selection system, and made the following findings.

Strengths of TM record selection systems

While the TM record selection systems were able to handle nouns and verbs with similar accuracy, the MT systems were much less capable of handling nouns. In rule-based MT systems, it is relatively easy to devise rules for verb sense disambiguation based on predicate-argument structures, but rules for noun sense disambiguation are generally harder to make as it is not clear what features to use. In the case of the TM, the human annotators have provided sufficient context around each word to allow for disambiguation, making the performance over

nouns comparable to that over verbs. Instances which were analyzed correctly by the two TM systems but no MT system contain such cases as the following:

- 記念 事業 commemorative events/project
  The MT systems mis-translated it as "commemoration enterprise" and "commemoration business". The TM systems handled it easily since it matched exactly with a TM record: 記念事業/commemorative project.
- 大学 時代 University days
  The MT systems mis-translated it as "era" and "age". The TM systems could handle it correctly using the TM record: 中学時代/junior high school days.
- 国内 に持ち込む bring/smuggle (something) into the country
  The MT systems mis-translated it as "at home" and "within the country". The TM systems matched it with the TM record: ··· を国内に持ち込む/to smuggle (something) into the country.
- ···場合 もあれば there are some case that ···/in some cases ···
  The MT systems mis-interpreted it as an "if" clause. The TM systems could handle it correctly based on the TM record: 体罰が必要な場合もある/Corporal punishment is needed in some cases.

Strengths of MT systems

For popular terms in politics and economics, the commercial MT systems tend to have a wide-coverage dictionary including the following examples, not contained in the TM.

- 一般競走入札 open bid system
- 一般教書演説 State of the Union address

With verbs, the biggest problem for TM selection systems is that the TM contains no information about the default (the most frequently used) record, and all records look equally important. In most cases, the default translation is used in a wider semantic domain, but the TM contains only a few records for that translation. That is, a few default records have to defend their wide semantic domain against many other specific records, putting the default records at a great disadvantage.

- 車を与える give a car
  The default record, 金を与える/to give money, lost out over more specific records such as "assign", "award", and "have a bad influence".
- 水を使う use water
  The default record, ナイフを使う/to use a knife, lost out over more specific records

such as “spend”, “waste”, and “be careful” (an idiom whose literal translation is “use a mind”).

- ジェットコースターに乗る ride the roller coaster
  The default record, 自転車に乗る/to ride on a bicycle, was not chosen due to competition from more specific records such as “catch” and “be well underway” (idiom whose literal translation is “ride an orbit”).

To solve these problems (low coverage and lack of defaults), we need to be able to auto-detect new translation patterns from bilingual corpora and automatically augment the TM from a variety of resources. The CRL-NYU system and the Ibaraki system extended the TM in these directions, but because of the tight time schedule of the exercise, extensions to the data were insufficient. Further investigation of the task results and efficient use of the gold standard test data may suggest means for successful semi/fully-automatic extension of the TM.

5 Conclusion

This paper has described the SENSEVAL-2 Japanese translation task: the task design, data construction, and task results. We think the task was successful since we were able to organize a challenging task, had a large number of participants, yielded reasonable results and promoted dialogue among the participants. The primary motivation for the task was well satisfied in that we escaped from the frustrating task of defining monolingual senses. The application task, machine translation, is very difficult, but challenging and interesting.

We have to increase the coverage of the TM and add more information such as frequency and what the default is. Furthermore, in order to construct a real MT system based on the TM, structural mappings between Japanese and English expressions are necessary. After solving those problems, we hope to have an exercise for TM-based machine translation systems in the near future.

Acknowledgement

We wish to express my gratitude to Mainichi Newspapers for providing articles. We would also like to thank Prof. Takenobu Tokunaga (Tokyo Institute of Technology) and Prof. Kiyoaki Shirai (JAIST) for their valuable advise about task organization, Yuiko Igura (Kyoto Univ.) and Inter Group Corp. for data construction, and all participants to the task.

Reference

Baldwin, T., Okazaki, A., Tokunaga, T., and Tanaka, H. (2001). "The Japanese Translation Task: Lexical and Structural Perspectives." In *Proceedings of SENSEVAL-2*, pp. 55–58.

Ide, N. (2000). "Cross-Lingual Sense Determination: Can It Work?." *Computers and the Humanities*, *34*(1), pp. 223–234.

Kilgarriff, A., and Palmer, M. (2000). "Introduction to the Special Issue on SENSEVAL." *Computers and the Humanities*, *34*(1), pp. 1–13.

Kumano, T., Kashioka, H., and Tanaka, H. (2003). "Japanese-English Translation Selection Using Vector Space Model (in Japanese)." *Journal of Natural Language Processing*, *10*(3), pp. 39–59.

Nagao, M. (1981). "A Framework of Mechanical Translation between Japanease and English by Analogy Priciple." In *Proc. of the International NATO Symposium on Artificial and Human Intelligence*.

Pinkham, J., Och, F. J., and Knight, K. (Eds.). (2001). *Data-driven Machine Translation*. Association for Computational Linguistics.

Shinnou, H. (2001). "Ibaraki: learning system of WSD rules developed for SENSEVAL-2 Japanese Translation Task (in Japanese)." In *Technical report of IEICE. NLC 2001-39*, pp. 25–38.

Shirai, K. (2003). "SENSEVAL-2 Japanese Dictionary Task (in Japanese)." *Journal of Natural Language Processing*, *10*(3), pp. 3–24.

Uchimoto, K., Sekine, S., Murata, M., and Isahara, H. (2003). "Word Translation by Combining an Example-Based Method and Machine Learning Models (in Japanese)." *Journal of Natural Language Processing*, *10*(3), pp. 87–114.

Sadao Kurohashi: Sadao Kurohashi received the B.S., M.S., and PhD in Electrical Engineering from Kyoto University in 1989, 1991 and 1994, respectively. He has been a visiting researcher of IRCS, University of Pennsylvania in 1994. He is currently an associate professor of the Graduate School of Information Science and Technology at the University of Tokyo. His research interests include natural language processing, knowledge acquisition/representation, and information retrieval.

Kiyotaka Uchimoto: Kiyotaka Uchimoto received the B.S. and M.S. in Electrical Engineering from Kyoto University in 1994 and 1996, respec-

tively. He is currently a research scientist of the Communications Research Laboratory, Japan. He is a member of the Association for Natural Language Processing, the Information Processing Society of Japan, and the Association for Computational Linguistics. His research interests include natural language processing and information retrieval.

(Received April 30, 2002)
(Revised August 8, 2002)
(Accepted August 18, 2002)