SENSEVAL-2 Japanese Translation Task

Sadao Kurohashi
University of Tokyo
kuro@kc.t.u-tokyo.ac.jp

Abstract

This paper reports an overview of SENSEVAL-2 Japanese translation task. In this task, word senses are defined according to translation distinction. 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 target word instance, a TM record best approximating that usage had to be submitted. Alternatively, submission could take the form of actual target word translations. 9 systems from 7 organizations participated in the task.

1 Introduction

In written texts, words which have multiple senses can be classified into two categories; homonyms and polysemous words. Generally speaking, while homonymy sense distinction is quite clear, polysemy sense distinction is very subtle and hard. English texts contain many homonyms. On the other hand, Japanese texts in which most content words are written by ideograms rarely contain homonyms. That is, the main target in Japanese WSD is polysemy, which makes Japanese WSD task setup very hard. What sense distinction of polysemous words is reasonable and effective heavily depends on how to use it, that is, an application of WSD.

Considering such a situation, in addition to the ordinary dictionary task we organized another task for Japanese, a translation task, in which word sense is defined according to translation distinction. Here, we set up the task assuming the example-based machine translation paradigm (Nagao, 1981). 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). We call a pair of Japanese and English expressions in the TM as a TM record. Given an evaluation document containing a target word, participants have 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 we can compare TM based systems with existing MT systems.

For evaluation, we distributed 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.

2 Construction of Translation Memory

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

1. By referring to the KWIC (Key Word In Context) of a target word, its typical Japanese expressions are picked up by lexicographers.
2. The Japanese expressions are translated by a translation company.

KWIC was made from the nine years volume of Mainichi Newspaper corpus. They are morphologically analyzed and segmented into phrase sequences, and then the 100 most frequent phrase uni-grams, bi-grams (two types; the target word is in the first phrase or the second phrase) and tri-grams (the target word is in the middle phrase) are provided to lexicographers (Figure 2).

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
...	...

Figure 1: An example of Translation Memory.

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。
...	...	...	...

Figure 2: An example of KWIC (numbers indicate phrase frequency).

The lexicographers pick up a typical expression of the target word from the KWIC. If its sense is context-independently clear, the expression is adopted as it is. If its sense is not clear, some pre/post expressions are supplemented by referring original sentences in the newspaper corpus.

Then, 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 records on average). The average number of words of a Japanese expression is 4.5.

3 Gold Standard Data and the Evaluation of Translations

As a gold standard data of the task, 40 target words were chosen out of 320 TM words. Considering the possible comparison of the translation task and the dictionary task, 40 target words were fully overlapped with 100 target words of the dictionary task.

In the Japanese dictionary task, target words are classified into three categories according to the difficulty (difficult, intermediate, easy), based on the entropy of word sense distribution in the training data of the dictionary task(Shirai, 2001). 40 target words of the trans lation task consists of 20 nouns and 20 verbs: difficult nouns and verbs, 10 intermediate nou and verbs, and 5 easy nouns and verbs.

For each target word, 30 instances were ch sen from Mainichi Newspaper corpus (in tot 1,200 instances) and they are also overlappe with the dictionary task. Since the dictiona task uses 100 instances for each target wor the translation task used 1st, 4th, 7th, ... 90t instances of the dictionary task.

As a gold standard data, zero or more a propriate TM records were assigned to each i stance by the same translation company. A propriate TM records were classified into th following three classes:

◎ : A TM record which can be used t translate the instance. POS, tense, plura singular, and subtle nuance do not nece sarily match.

○ : If the instance is considered alone, th English translation is correct, but usin the TM record in the given context is nc so good, for example, making very round about translation.

△ : If the instance is considered alone, the English translation is correct, but using the TM record in the given context is inappropriate.

Out of 1,200 instances, 34 instances (2.8%) were assigned no TM records (there was no appropriate TM record). To one instance, on average, 6.6 records were assigned as ◎, 1.4 records as ○, and 0.1 records as △, in total 8.1 records. If a system chooses a TM record randomly as an answer, the accuracy becomes 36.8% in case that all of ◎, ○ and △ records are regarded as correct, and 29.0% in case that only ◎ is regarded as correct (they are the baseline scores used in the next section).

In the gold standard data construction, 90 instances (9 words × 10 instances) were dealt with by two annotators doubly, and then their agreement were checked. For each instance one record is chosen randomly from annotator B's answers, and it was checked whether it is contained in annotator A's answers (annotator A made the whole gold standard data). The agreement was 86.6% in case that all of ◎, ○ and △ records are regarded as correct, and 80.9% in case that only ◎ is regarded as correct.

In the case that the submission is in the form of translation data, translation experts (the same company as constructed the TM and the gold standard data) were asked to rank the supplied translation ◎, ○ or ×. This evaluation does not pay attention to the total translation, but just the appropriateness of the target instance translation.

4 Result

In the Japanese translation task, 9 systems from 7 organizations submitted the answers. The characteristics of the systems are summarized as follows:

- AnonymX, AnonymY
  Commercial, rule-based MT systems.
- CRL-NYU (Communications Research Laboratory & New York Univ.)
  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.
- Ibaraki (Ibaraki Univ.)
  A training data was constructed manually from newspaper articles, 170 instances for each target word. Features were collected in 7-word window around the target word, and decision list method was used for learning.
- Stanford-Titech1 (Stanford Univ. & Tokyo Institute of Technology)
  The system selects the appropriate TM record based on the character-bigram-based Dice's coefficient. It also utilized the context of the other target word instances in the evaluation text.
- AnonymZ
  A sentence (TM records for learning, and an input for testing) is morphologically analyzed and converted into a semantic tag sequence, and maximum entropy method was used for learning.
- ATR
  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 bottom-up, shared-memory based matching algorithm.
- Stanford-Titech2 (Stanford Univ. & Tokyo Institute of Technology)
  The system selects the appropriate TM record based on the case-frame-based similarity, using NTT Goi-Taikei thesaurus.

The results of all systems are shown in Figure 3. The left bar charts indicate the accuracy based on the lenient evaluation (◎, ○ and △ in TM selection and ◎ and ○ in MT are regarded as correct); the right bar charts indicate the accuracy based on the strict evaluation (◎ is only regarded as correct both in TM selection and MT). Note that since the TM does not have

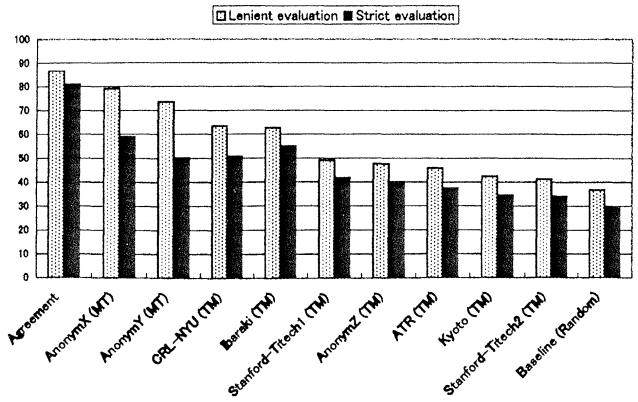

Figure 3: Result of the Japanese translation task.

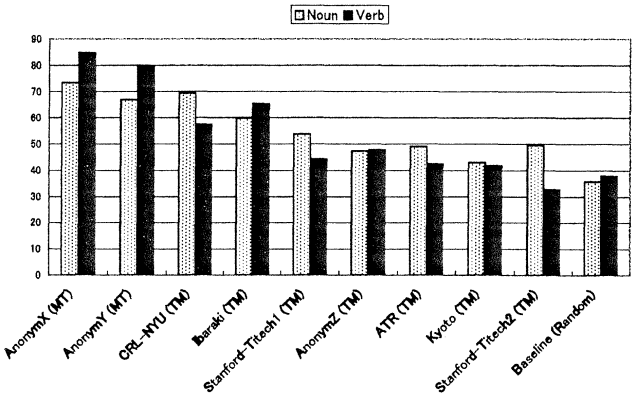

Figure 4: Scores for nouns and verbs.

a hierarchical structure, there is no evaluation options such as fine, coarse, and mixed.

Figure 4 shows scores for nouns and verbs separately, and Figure 5 shows scores for difficult/intermediate/easy words. Both of them were evaluated by the lenient criteria.

In these figures, "Agreement" and "Baseline" were as described in the previous section. When the system judges that there is no appropriate TM record for an instance, it can return "UNASSIGNABLE". In that case, if there is no appropriate TM record assigned in the gold standard data, the answer is regarded as correct.

Among TM selection systems, systems using some extra learning data outperformed other systems just using the TM. The comparison between TM selection systems and MT systems is not easy, but the result indicates the effectiveness of the accumulated know-how of MT systems. However, the performance of the best TM selection system is not so different from MT

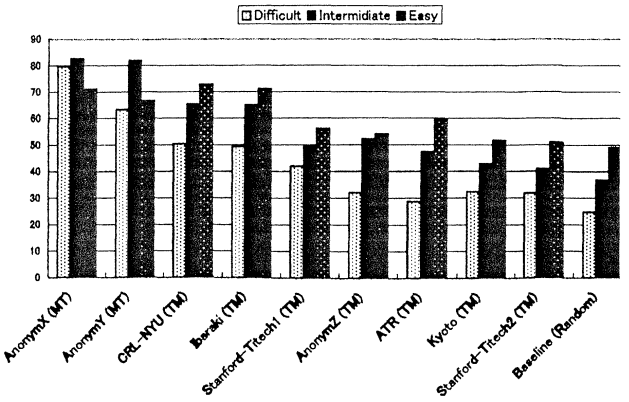

Figure 5: Scores for difficulty classes.

systems, which indicates the promising future of TM based techniques.

5 Conclusion

This paper described an overview of SENSEVAL-2 Japanese translation task. The data used in this task are available at SENSEVAL-2 web site. We hope this valuable data helps improve WSD and MT systems.

Acknowledgment

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

References

Makoto Nagao. 1981. A framework of mechanical translation between Japanease and English by analogy priciple. In *Proc. of the International NATO Symposium on Artificia and Human Intelligence.*

Kiyoaki Shirai. 2001. SENSEVAL-2 Japanese dictionary task. In *Proceedings of the SENSEVAL-2 Workshop.*