SENSEVAL-2 Japanese Dictionary Task

Kiyoaki Shirai
School of Information Science, Japan Advanced Institute of Science and Technology
kshirai@jaist.ac.jp

Abstract

This paper reports an overview of the SENSEVAL-2 Japanese dictionary task. It was a lexical sample task, and word senses are defined according to a Japanese dictionary, the Iwanami Kokugo Jiten. The Iwanami Kokugo Jiten and a training corpus were distributed to all participants. The number of target words was 100, 50 nouns and 50 verbs. One hundred instances of each target word were provided, making for a total of 10,000 instances for evaluation. Seven systems of three organizations participated in this task.

1 Introduction

In SENSEVAL-2, there are two Japanese tasks, a translation task and a dictionary task. This paper describes the details of the dictionary task.

First of all, let me introduce an overview of the Japanese dictionary task. This task is a lexical sample task. Word senses were defined according to the Iwanami Kokugo Jiten (Nishio et al., 1994), a Japanese dictionary published by Iwanami Shoten. It was distributed to all participants as a sense inventory. Training data, a corpus consisting of 3,000 newspaper articles and manually annotated with sense IDs, was also distributed to participants. For evaluation, we distributed newspaper articles with marked target words as test documents. Participants were required to assign one or more sense IDs to each target word, optionally with associated probabilities. The number of target words was 100, 50 nouns and 50 verbs. One hundred instances of each target word were provided, making for a total of 10,000 instances.

In what follows, Section 2 describes details of data used in the Japanese dictionary task. Section 3 describes the process to construct the gold standard data, including the analysis of inter-tagger agreement. Section 4 briefly introduces participating systems and their results. Finally, Section 5 concludes this paper.

2 Data

In the Japanese dictionary task, three data were distributed to all participants: sense inventory, training data and evaluation data.

2.1 Sense Inventory

As described in Section 1, word senses are defined according to a Japanese dictionary, the Iwanami Kokugo Jiten. The number of headwords and word senses in the Iwanami Kokugo Jiten is 60,321 and 85,870, respectively.

Figure 1 shows an example of word sense descriptions in the Iwanami Kokugo Jiten, the sense set of the Japanese noun "MURI."

MURI
1. lack of reasonableness
   1-a. something not to be rational, not to be sensible [*kimi ga okoru no wa* ***MURI*** *mo nai* (It is natural for you to be angry)]
   1-b. to do something compulsorily [*sigoto no* ***MURI*** *de byouki ni naru* (I become ill from overwork)]

Figure 1: Sense set of "MURI"

As shown in Figure 1, there are hierarchical structures in word sense descriptions. For example, word sense 1 subsumes 1-a and 1-b. The number of layers of hierarchy in the Iwanami Kokugo Jiten is at most 3. Word sense distinctions in the lowest level are rather fine or subtle. Furthermore, a word sense description sometimes contains example sentences including a headword, indicated by italics in Figure 1.

The Iwanami Kokugo Jiten was provided to all participants. For each sense description, a

corresponding sense ID and morphological information were supplied. All morphological information, which included word segmentation, part-of-speech (POS) tag, base form and reading, was manually post-edited.

2.2 Training Data

An annotated corpus was distributed as the training data. It was made up of 3,000 newspaper articles extracted from the 1994 Mainichi Shimbun, consisting of 888,000 words. The annotated information in the training corpus was as follows:

- Morphological information

  The text was annotated with morphological information (word segmentation, POS tag, base form and reading) for all words. All morphological information was manually post-edited.

- UDC code

  Each article was assigned a code representing the text class. The classification code system was the third version (INFOSTA, 1994) of Universal Decimal Classification (UDC) code (Organization, 1993).

- Word sense IDs

  Only 148,558 words in the text were annotated for sense. Words assigned with sense IDs satisfied the following conditions:

  1. Their POSs were noun, verb or adjective.
  2. The Iwanami Kokugo Jiten gave sense descriptions for them.
  3. They were ambiguous, i.e. there are more than two word senses in the dictionary.

  Word sense IDs were manually annotated. However, only one annotator assigned a sense ID for each word.

2.3 Evaluation Data

The evaluation data was made up of 2,130 newspaper articles extracted from the 1994 Mainichi Shimbun. The articles used for the training and evaluation data were mutually exclusive. The annotated information in the evaluation data was as follows:

- Morphological information

  The text was annotated with morphological information (word segmentation, POS tag, base form and reading) for all words. Note that morphological information in the training data was manually post-edited, but not in the evaluation data. So participants might ignore morphological information in the evaluation data.

- UDC code

  As in the training data. each article was assigned a UDC code

- Word sense IDs (gold standard data)

  Word sense IDs were annotated manually for the target words [1]. Note that word sense IDs in the evaluation and training data were given in different ways: (1) a sense ID was assigned for each word by at least two annotators in the evaluation data, while by only one annotator in the training data, (2) only 10,000 instances in the articles were annotated with sense IDs in the evaluation data, while all words were annotated which satisfied the conditions described in 2.2 in the training data.

3 Gold Standard Data

Except for the gold standard data, the data described in Section 2 have been developed by Real World Computing Partnership (Hasida et al., 1998; Shirai et al., 2001) and already released to public domain [2]. On the other hand, the gold standard data was newly developed for the SENSEVAL-2. This section presents the process of preparing the gold standard data, and the analysis of inter-tagger agreement.

3.1 Sampling Target Words

When we chose target words, we considered the following:

- POSs of target words were either nouns or verbs.
- Words were chosen which occurred more than 50 times in the training data.

[1] They were hidden from participants at the contest.

[2] Notice that the training data had been released to the public before the contest began. This violated the SENSEVAL-2 schedule constraint that answer submission should not occur more than 21 days after downloading the training data.

Table 1: Number of Target Words

	D_a	D_b	D_c	all
nouns	**10** (9.1/1.19)	**20** (3.7/0.723)	**20** (3.3/0.248)	**50** (4.6/0.627)
verbs	**10** (18/1.77)	**20** (6.7/0.728)	**20** (5.2/0.244)	**50** (8.3/0.743)
all	**20** (14/1.48)	**40** (5.2/0.725)	**40** (4.2/0.246)	**100** (6.5/0.685)

(average polysemy / average entropy)

- The relative "difficulty" in disambiguating the sense of words was considered. Difficulty of the word w was defined by the entropy of the word sense distribution $E(w)$ in the training data. Obviously, the higher $E(w)$ was, the more difficult the WSD for w was.

  We set up three word classes, D_a ($E(w) \geq 1$), D_b ($0.5 \leq E(w) < 1$) and D_c ($E(w) < 0.5$), and chose target words evenly from them.

Table 1 reveals details of numbers of target words. Average polysemy (i.e. average number of word senses per headword) and average entropy are also indicated.

One hundred instances of each target word were selected from newspaper articles, making for a total of 10,000 instances.

3.2 Manual Annotation

Six annotators assigned the correct word sense IDs for 10,000 instances. They were not experts, but had knowledge of linguistics or lexicography to some degree. The process of manual annotating was as follows:

Step 1. Two annotators chose a sense ID for each instance separately in accordance with the following guidelines:

- Only one sense ID was to be chosen for each instance.
- Sense IDs at any layers in hierarchical structures could be assignable.
- The "UNASSIGNABLE" tag was to be chosen only when all sense IDs weren't absolutely applicable. Otherwise, choose one of sense IDs in the dictionary.

Table 2: Inter-tagger Agreement

	D_a	D_b	D_c	(all)
nouns	0.809	0.786	0.957	0.859
verbs	0.699	0.896	0.922	0.867
all	0.754	0.841	0.939	0.863

Step 2. If the sense IDs selected by 2 annotators agreed, we considered it to be a correct sense ID for an instance.

Step 3. If they did not agree, the third annotator chose the correct sense ID between them. If the third annotator judged both of them to be wrong and chose another sense ID as correct, we considered that all 3 word sense IDs were correct.

According to Step 3., the number of words for which 3 annotators assigned different sense IDs from one another was a quite few, 28 (0.3%).

Table 2 indicates the inter-tagger agreement of two annotators in Step 1. Agreement ratio for all 10,000 instances was 86.3%.

4 Results for Participating Systems

In the Japanese dictionary task, the following 7 systems of 3 organizations submitted answers. Notice that all systems used supervised learning techniques.

- Communications Research Laboratory and New York University (CRL1 $\sim$ CRL4)

  The learning schemes were simple Bayes and support vector machine (SVM), and two kinds of hybrid models of simple Bayes and SVM.

- Tokyo Institute of Technology (Titech1, Titech2)

  Decision lists were learned from the training data. The features used in the decision lists were content words and POS tags in a window, and content words in example sentences contained in word sense descriptions in the Iwanami Kokugo Jiten.

- Nara Institute of Science and Technology (Naist)

  The learning algorithm was SVM. The feature space was reconstructed using Principle Component Analysis(PCA) and Independent Component Analysis(ICA).

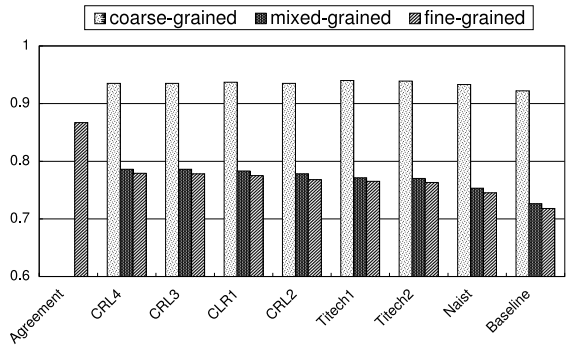

Figure 2: Results

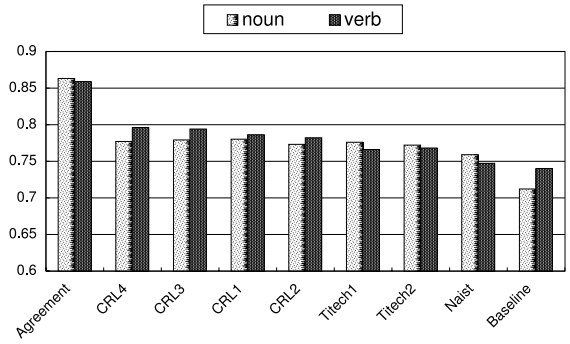

Figure 3: Mixed-grained scores for nouns and verbs

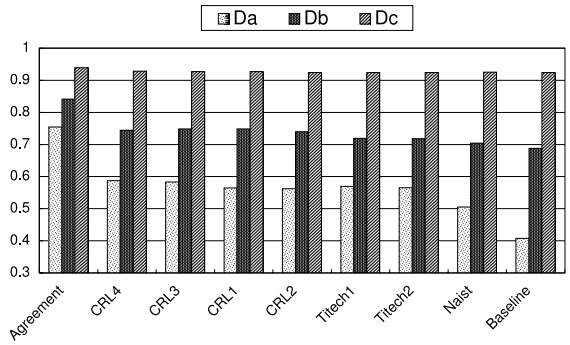

Figure 4: Mixed-grained scores for word classes

The results of all systems are shown in Figure 2. "Baseline" indicates the system which always selects the most frequent word sense ID, while "Agreement" indicates the agreement ratio between two annotators. All systems outperformed the baseline, and there was no remarkable difference between their scores (differences were 3 % at most).

Figure 3 indicates the mixed-grained scores for nouns and verbs. Comparing baseline system scores, the score for verbs was greater than that for nouns, even though the average entropy of verbs was higher than that of nouns (Table 1). The situation was the same in CRL systems, but not in Titech and Naist. The reason why the average entropy was not coincident with the score of the baseline was that the entropy of some verbs was so great that it raised the average entropy disproportionately. Actually, the entropy of 7 verbs was greater than the maximum entropy of nouns.

Figure 4 indicates the mixed-grained scores for each word class. For word class D_c, there was hardly any difference among scores of all systems, including Baseline system and Agreement. On the other hand, appreciable difference was found for D_a and D_b.

5 Conclusion

This paper reports an overview of the SENSEVAL-2 Japanese dictionary task. The data used in this task are available on the SENSEVAL-2 web site. I hope this valuable data helps all researchers to improve their WSD 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. Sadao Kurohashi (University of Tokyo) for valuable advise about task organization, the annotators for constructing gold standard data, and all participants.

References

Koiti Hasida et al. 1998. The RWC text databases. In *Proceedings of the the first International Conference on Language Resources and Evaluation*, pages 457–462.

INFOSTA. 1994. *Universal Decimal Classification*. Maruzen, Tokyo. (in Japanese).

Minoru Nishio, Etsutaro Iwabuchi, and Shizuo Mizutani. 1994. *Iwanami Kokugo Jiten Dai Go Han*. Iwanami Publisher. (in Japanese).

British Standards Organization. 1993. *Guide to the Universal Decimal Classification (UDC)*. BSI, London.

Kiyoaki Shirai et al. 2001. Text database with word sense tags defined by Iwanami Japanese dictionary. *SIG notes of Information Processing Society of Japan*, 2001(9):117–122. (in Japanese).