Construction of a Word Sense Tagged Corpus for SENSEVAL-2 Japanese Dictionary Task

Kiyoaki Shirai*

*Japan Advanced Institute of Science and Technology
1-1 Asahidai, Tatsunokuchi, Ishikawa, Japan
kshirai@jaist.ac.jp

Abstract

This paper reports the details of a Japanese word sense tagged corpus developed as an evaluation data for SENSEVAL-2 Japanese dictionary task. The corpus made up of 2,130 newspaper articles. Not all but only 10,000 words in the articles were manually annotated with sense IDs, which was used as a gold standard data. Word senses were defined according to the Iwanami Kokugo Jiten, a Japanese dictionary published by Iwanami Shoten. Two annotators chose a sense ID for each instance separately. If they did not agree, the third annotator chose the correct sense ID between them. Inter-tagger agreement and Cohen's κ was 86.3% and 0.677, respectively.

1. Introduction

This paper reports the process to construct a Japanese word sense tagged corpus, and statistical analysis such as inter-tagger agreement, Cohen's κ etc. The corpus described here was developed as a evaluation data of SENSEVAL-2 Japanese dictionary task (Shirai, 2001). SENSEVAL (Kilgarriff and Palmer, 2000) is an evaluation exercise for word sense disambiguation.

In Section 2., SENSEVAL-2 Japanese dictionary task is briefly introduced. The process to construct the word sense tagged corpus is described in Section 3., while a statistical analysis in Section 4. Finally, I conclude the paper in Section 5.

2. SENSEVAL-2 Japanese dictionary task

SENSEVAL-2 was the evaluation exercise for word sense disambiguation (WSD hereafter), held in the spring of 2001. 12 languages including Japanese were represented in the competition. This section gives a brief introduction of Japanese dictionary task.

Japanese dictionary task was 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. Each sense ID corresponded to one of word sense descriptions in the Iwanami Kokugo Jiten. 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. 7 systems of 3 organizations participated in this task.

Real World Computing Partnership (RWCP) have already developed a word sense tagged corpus in 1998-2000 (Shirai et al., 2001), and we used it as a training data in SENSEVAL-2. As it has released to public domain before the competition began, a word sense tagged corpus hidden from participants was newly developed early in 2001 as an evaluation data. This paper focuses on this newly developed corpus and describes the details of it.

3. Construction of an Evaluation Data

3.1. Text

The evaluation data was the corpus 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 corpus was annotated with morphological information, i.e. word segmentation, POS tag, base form and reading for all words. Furthermore, 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 (British Standards Organization, 1993). All morphological information was automatically annotated, while UDC codes manually. This corpus has been developed by RWCP (Hasida et al., 1998).

Word sense IDs were newly assigned to this corpus to construct a gold standard data for SENSEVAL-2 Japanese dictionary task.

3.2. Sense Inventory

As described in Section 2., word sense IDs were 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. Therefore, average polysemy, i.e. average number of word senses per a headword, is 1.42.

Figure 1 shows an example of word sense descriptions in the Iwanami Kokugo Jiten, the sense set of the Japanese noun "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. A word sense description sometimes contains example sentences including a headword, indicated by italics in Figure 1. These example sentences would also help annotators to select correct word senses.

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"

	D_a	D_b	D_c	all
nouns	10	20	20	50
verbs	10	20	20	50
all	20	40	40	100

Table 1: Number of Target Words

	D_a	D_b	D_c	all
nouns	9.1	3.7	3.3	4.6
verbs	18	6.7	5.2	8.3
all	14	5.2	4.2	6.4

Table 2: Average Polysemy of Target Words

	D_a	D_b	D_c	all
nouns	1.19	0.723	0.248	0.627
verbs	1.77	0.728	0.244	0.743
all	1.48	0.725	0.246	0.685

Table 3: Average Entropy of Target Words

3.3. Sampling Target Words

Japanese dictionary task was a lexical sample task, i.e. participating systems had to disambiguate not all words but only sample words in the text. The number of target words was set to be 100. 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.
- 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 following three word classes and chose target words evenly from them.

  $$\begin{array}{ll} D_a & E(w) \geq 1 \\ D_b & 0.5 \leq E(w) < 1 \\ D_c & E(w) < 0.5 \end{array}$$

Table 1, 2 and 3 reveals details of numbers of target words, average polysemy and average entropy in the training data, respectively.

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

3.4. 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.

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%).

As described in Section 2., we used an existing word sense tagged corpus as a training data, while developed another word sense tagged corpus for an evaluation data. Note that word sense IDs in the evaluation and training data were given in different ways:

- 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.
- only 10,000 instances in the articles were annotated with sense IDs in the evaluation data, while all words which satisfied the following conditions were annotated in the training data:
  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.

	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

Table 4: Inter-tagger Agreement

	D_a	D_b	D_c	all
nouns	0.705	0.531	0.775	0.659
verbs	0.593	0.719	0.721	0.694
all	0.649	0.630	0.746	0.677

Table 5: Cohen's κ

4. Statistics

Table 4 indicates the inter-tagger agreement of two annotators. Agreement ratio for all 10,000 instances was 86.3%. According to POS, agreement ratios for nouns and verbs were almost equal: 85.9% and 86.7%, respectively. On the other hand, the more difficult WSD was, the lower agreement ratio became (notice that the order of difficulty is $D_a > D_b > D_c$).

Table 4 shows averages of Cohen's κ for a target word. κ is a statistical measure indicating a degree of agreement of annotators, and evaluates the observed proportion of agreement (P_o) against the expected proportion of agreement (P_e) as shown in (1).

$$\kappa = \frac{P_o - P_e}{1 - P_e} \quad (1)$$

$$P_o = \frac{\sum\sum n_{ij}^2 - Nk}{Nk(k-1)} \quad (2)$$

$$P_e = \sum_{j=1}^{v} \left(\frac{\sum_{i=1}^{N} n_{ij}}{Nk} \right)^2 \quad (3)$$

In (2) and (3), N denotes the number of instances, k the number of annotators, v the number of different word sense IDs that can be assigned to instances, and n_{ij} the number of annotators assigning instance i to word sense ID j. κ shown in Table 5 was relatively low. Average of κ for all 100 words was 0.677.

Figure 2 represents relation between inter-tagger agreement and system's score. In each graph, the horizontal axis indicates the inter-tagger agreement of each word, while the vertical axis indicates the average of mix-grained score[1] of participant's 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 (4 systems)

  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 (2 systems)

  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 (1 system)

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

According to Figure 2, an inter-tagger agreement and an average of system's scores was positively correlated, i.e. the higher an inter-tagger agreement was, the higher system's scores were. However, system's scores for some words, such as "me"(eyes), "$jikan$"(time), "$kaihatsu$" (development), "$kakaru$"(hang, require etc.), "$kiku$"(hear) and "$noru$"(ride), were low even though inter-tagger agreement were high. In other words, systems could not consider effective features, which human annotators might consider, to disambiguate meanings of these words. Analysis on the reason why systems failed to select correct word senses for these words would give fruitful information to improve performances of WSD systems.

5. Conclusion

This paper reports the details of the process of preparing the word sense tagged corpus used as the gold standard data for SENSEVAL-2 Japanese task and the statistical analysis of it. It can be used for an evaluation of a Japanese WSD system, development of WSD systems, etc.

The annotation data is freely available at SENSEVAL-2 web site (SENSEVAL-2, 2001). However, the text is not public due to copyright restrictions. Because we use text excerpts from 1994 Mainichi Shimbun newspaper articles as described in Section 3.1. If you purchase newspaper articles from the newspaper company, you can reconstruct the complete word sense tagged corpus. Purchase details as well as reconstruction tools are also available at the same web site.

6. References

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

Koiti Hasida, Hitoshi Isahara, Takenobu Tokunaga, Minako Hashimoto, Shiho Ogino, Wakako Kashino, Jun Toyoura, and Hironobu Takahashi. 1998. The RWC text databases. In *Proceedings of the first International Conference on Language Resources and Evaluation*, pages 457–462.

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

Adam Kilgarriff and Martha Palmer. 2000. Special issue on SENSEVAL: Evaluating word sense disambiguation programs. *Computers and the Humanities*, 34(1-2).

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

[1] "Mix-grained score" is one of the scoring scheme of SENSEVAL-2. See (SENSEVAL-2, 2001) for details.

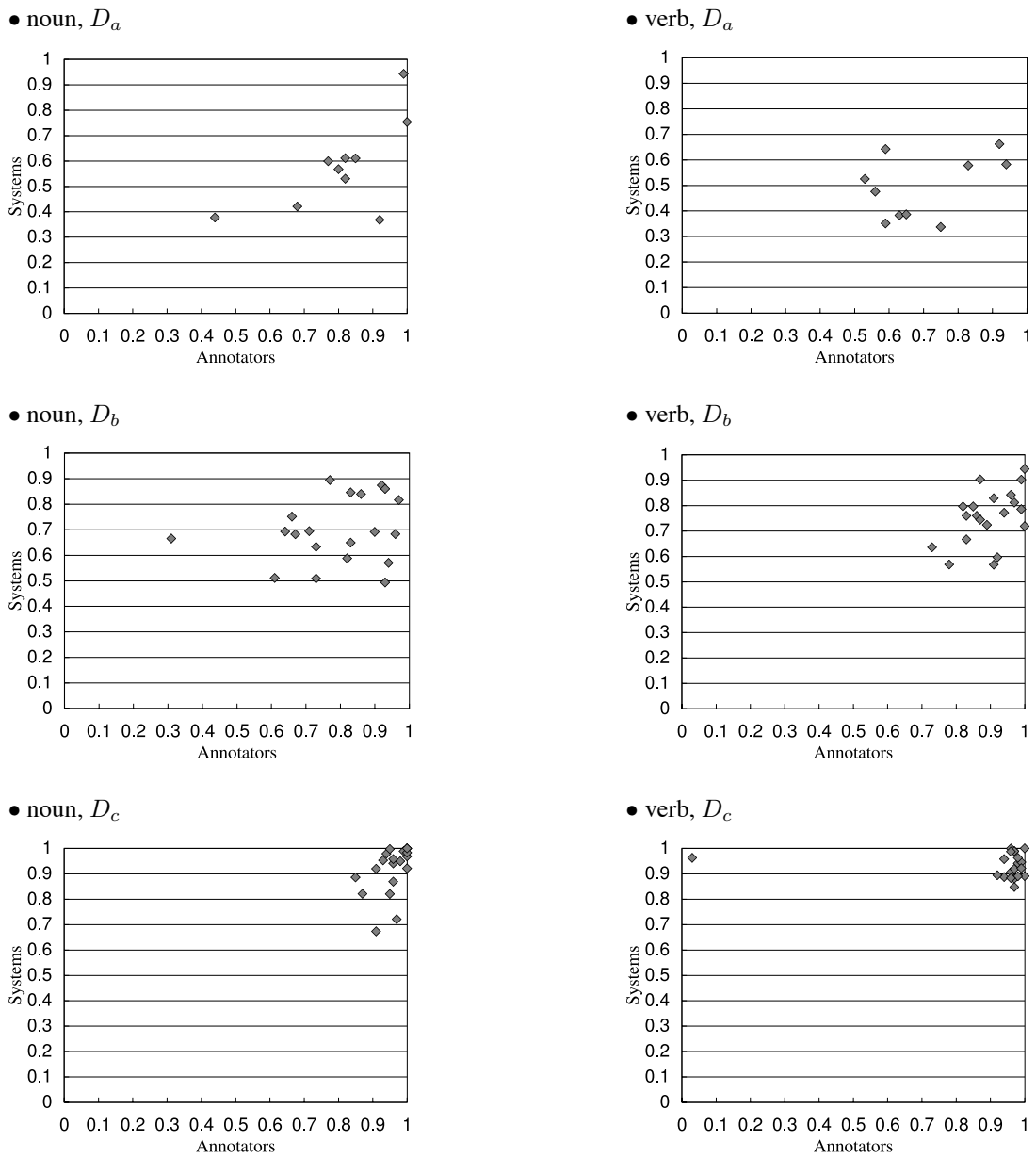

Figure 2: Relation between Inter-tagger Agreement and Performance of Participant's Systems

SENSEVAL-2. 2001. SENSEVAL-2 web site. `http://www.sle.sharp.co.uk/senseval2/`.

Kiyoaki Shirai, Wakako Kashino, Minako Hashimoto, Takenobu Tokunaga, Eiichi Arita, Hitoshi Isahara, Shiho Ogino, Ryuichi Kobune, Hironobu Takahashi, Katashi Nagao, Hasida Kôiti, and Murata Masaki. 2001. Text database with word sense tags defined by Iwanami Japanese dictionary. In *IPSJ SIG Notes, NL-141*, pages 117–122. (in Japanese).

Kiyoaki Shirai. 2001. SENSEVAL-2 Japanese dictionary task. In *Proceedings of the SENSEVAL-2: Second International Workshop on Evaluating Word Sense Diasambiguating Systems*. (to appear).