SemEval-2010 Task: Japanese WSD

Manabu Okumura
Tokyo Institute of Technology
oku@pi.titech.ac.jp

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

Kanako Komiya
Tokyo University of Agriculture and Technology
kkomiya@cc.tuat.ac.jp

Hikaru Yokono
Tokyo Institute of Technology
yokono@lr.pi.titech.ac.jp

Abstract

An overview of the SemEval-2 Japanese WSD task is presented. It is a lexical sample task, and word senses are defined according to a Japanese dictionary, the Iwanami Kokugo Jiten. This dictionary and a training corpus were distributed to participants. The number of target words was 50, with 22 nouns, 23 verbs, and 5 adjectives. Fifty instances of each target word were provided, consisting of a total of 2,500 instances for the evaluation. Nine systems from four organizations participated in the task.

1 Introduction

This paper reports an overview of the SemEval-2 Japanese Word Sense Disambiguation (WSD) task. It can be considered an extension of the SENSEVAL-2 Japanese monolingual dictionary-based task (Shirai, 2001), so it is a lexical sample task. Word senses are defined according to the Iwanami Kokugo Jiten (Nishio et al., 1994), a Japanese dictionary published by Iwanami Shoten. It was distributed to participants as a sense inventory. Our task has the following two new characteristics:

1. All previous Japanese sense-tagged corpora were from newspaper articles, while sense-tagged corpora were constructed in English on balanced corpora, such as Brown corpus and BNC corpus. The first balanced corpus of contemporary written Japanese (BCCWJ corpus) is now being constructed as part of a national project in Japan (Maekawa, 2008), and we are now constructing a sense-tagged corpus based on it. Therefore, the task will use the first balanced Japanese sense-tagged corpus.

   Because a balanced corpus consists of documents from multiple genres, the corpus can be divided into multiple sub-corpora of a genre. In supervised learning approaches on word sense disambiguation, because word sense distribution might vary across different sub-corpora, we need to take into account the genres of training and test corpora. Therefore, word sense disambiguation on a balanced corpus requires tackling a kind of domain (genre) adaptation problem (Chang and Ng, 2006; Agirre and de Lacalle, 2008).

2. In previous WSD tasks, systems have been required to select a sense from a given set of senses in a dictionary for a word in one context (an instance). However, the set of senses in the dictionary is not always complete. New word senses sometimes appear after the dictionary has been compiled. Therefore, some instances might have a sense that cannot be found in the dictionary's set. The task will take into account not only the instances that have a sense in the given set but also the instances that have a sense that cannot be found in the set. In the latter case, systems should output that the instances have a sense that is not in the set.

Training data, a corpus that consists of three genres (books, newspaper articles, and white papers) and is manually annotated with sense IDs, was also distributed to participants. For the evaluation, we distributed a corpus that consists of four genres (books, newspaper articles, white papers, and documents from a Q&A site on the WWW) with marked target words as test data. Participants were requested to assign one or more sense IDs to each target word, optionally with associated probabilities. The number of target words was 50, with 22 nouns, 23 verbs, and 5 adjectives. Fifty instances of each target word were provided, con-

Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, pages 69–74,
Uppsala, Sweden, 15-16 July 2010.

sisting of a total of 2,500 instances for the evaluation.

In what follows, section two describes the details of the data used in the Japanese WSD task. Section three describes the process to construct the sense tagged data, including the analysis of an inter-annotator agreement. Section four briefly introduces participating systems and section five describes their results. Finally, section six concludes the paper.

2 Data

In the Japanese WSD task, three types of data were distributed to all participants: a sense inventory, training data, and test data[1].

2.1 Sense Inventory

As described in section one, 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.

As described in the task description of SENSEVAL-2 Japanese dictionary task (Shirai, 2001), the Iwanami Kokugo Jiten has hierarchical structures in word sense descriptions. The Iwanami Kokugo Jiten has at most three hierarchical layers.

2.2 Training Data

An annotated corpus was distributed as the training data. It consists of 240 documents of three genres (books, newspaper articles, and white papers) from the BCCWJ corpus. The annotated information in the training data is as follows:

- Morphological information
  The document was annotated with morphological information (word boundaries, a part-of-speech (POS) tag, a base form, and a reading) for all words. All the morphological information was automatically annotated using chasen[2] with unidic and was manually post-edited.

- Genre code
  Each document was assigned a code indicating its genre from the aforementioned list.

- Word sense IDs
  3,437 word types in the data were annotated for sense IDs, and the data contain 31,611 sense-tagged instances that include 2,500 instances for the 50 target words. Words assigned with sense IDs satisfied the following conditions:

  1. The Iwanami Kokugo Jiten gave their sense description.
  2. Their POSs were either a noun, a verb, or an adjective.
  3. They were ambiguous, that is, there were more than two word senses for them in the dictionary.

  Word sense IDs were manually annotated.

2.3 Test Data

The test data consists of 695 documents of four genres (books, newspaper articles, white papers, and documents from a Q&A site on the WWW) from the BCCWJ corpus, with marked target words. The documents used for the training and test data are not mutually exclusive. The number of overlapping documents between the training and test data is 185. The instances used for the evaluation were not provided as the training data[3]. The annotated information in the test data is as follows:

- Morphological information
  Similar to the training data, the document was annotated with morphological information (word boundaries, a POS tag, a base form, and a reading) for all words. All morphological information was automatically annotated using chasen with unidic and was manually post-edited.

- Genre code
  As in the training data, each document was assigned a code indicating its genre from the aforementioned list.

- Word sense IDs
  Word sense IDs were manually annotated for

[1] Due to space limits, we unfortunately cannot present the statistics of the training and test data, such as the number of instances in different genres, the number of instances for a new word sense, and the Jensen Shannon (JS) divergence (Lin, 1991; Dagan et al., 1997) between the word sense distributions of two different genres. We hope we will present them in another paper in the near future.

[2] `http://chasen-legacy.sourceforge.jp/`

[3] The word sense IDs for them were hidden from the participants.

the target words[4].

The number of target words was 50, with 22 nouns, 23 verbs, and 5 adjectives. Fifty instances of each target word were provided, consisting of a total of 2,500 instances for the evaluation.

3 Word Sense Tagging

Except for the word sense IDs, the data described in section two was developed by the National Institute of Japanese Language. However, the word sense IDs were newly annotated on the data. This section presents the process of annotating the word sense IDs, and the analysis of the inter-annotator agreement.

3.1 Sampling Target Words

When we chose target words, we considered the following conditions:

- The POSs of target words were either a noun, a verb, or an adjective.
- We chose words that occurred more than 50 times in the training data.
- The relative "difficulty" in disambiguating the sense of words was taken into account. The difficulty of the word *w* was defined by the entropy of the word sense distribution $E(w)$ in the test data (Kilgarriff and Rosenzweig, 2000). Obviously, the higher $E(w)$ is, the more difficult the WSD for *w* is.
- The number of instances for a new sense was also taken into account.

3.2 Manual Annotation

Nine annotators assigned the correct word sense IDs for the training and test data. All of them had a certain level of linguistic knowledge. The process of manual annotation was as follows:

1. An annotator chose a sense ID for each word separately in accordance with the following guidelines:
   - One sense ID was to be chosen for each word.
   - Sense IDs at any layers in the hierarchical structures were assignable.
   - The "new word sense" tag was to be chosen only when all sense IDs were not absolutely applicable.
2. For the instances that had a 'new word sense' tag, another annotator reexamined carefully whether those instances really had a new sense.

Because a fragment of the corpus was tagged by multiple annotators in a preliminary annotation, the inter-annotator agreement between the two annotators in step 1 was calculated with Kappa statistics. It was 0.678.

4 Evaluation Methodology

The evaluation was returned in the following two ways:

1. The outputted sense IDs were evaluated, assuming the 'new sense' as another sense ID. The outputted sense IDs were compared to the given gold standard word senses, and the usual precision measure for supervised word sense disambiguation systems was computed using the scorer. The Iwanami Kokugo Jiten has three levels for sense IDs, and we used the middle-level sense in the task. Therefore, the scoring in the task was 'middle-grained scoring.'
2. The ability of finding the instances of new senses was evaluated, assuming the task as classifying each instance into a 'known sense' or 'new sense' class. The outputted sense IDs (same as in 1.) were compared to the given gold standard word senses, and the usual accuracy for binary classification was computed, assuming all sense IDs in the dictionary were in the 'known sense' class.

5 Participating Systems

In the Japanese WSD task, 10 organizations registered for participation. However, only the nine systems from four organizations submitted the results. In what follows, we outline them with the following description:

1. learning algorithm used,
2. features used,
3. language resources used,

[4]They were hidden from the participants during the formal run.

4. level of analysis performed in the system,

5. whether and how the difference in the text genre was taken into account,

6. method to detect new senses of words, if any.

Note that most of the systems used supervised learning techniques.

- HIT-1
  1. Naive Bayes, 2. Word form/POS of the target word, word form/POS before or after the target word, content words in the context, classes in a thesaurus for those words in the context, the text genre, 3. 'Bunrui-Goi-Hyou', a Japanese thesaurus (National Institute of Japanese Language, 1964), 4. Morphological analysis, 5. A genre is included in the features. 6. Assuming that the posterior probability has a normal distribution, the system judges those instances deviating from the distribution at the 0.05 significance level as a new word sense

- JAIST-1
  1. Agglomerative clustering, 2. Bag-of-words in context, etc. 3. None, 4. Morphological analysis, 5. The system does not merge example sentences in different genre sub-corpus into a cluster. 6. First, the system makes clusters of example sentences, then measures the similarity between a cluster and a sense in the dictionary, finally regarding the cluster as a collection of new senses when the similarity is small. For WSD, the system chooses the most similar sense for each cluster, then it considers all the instances in the cluster to have that sense.

- JAIST-2
  1. SVM, 2. Word form/POS before or after the target word, content words in the context, etc. 3. None, 4. Morphological analysis, 5. The system was trained with the feature set where features are distinguished whether or not they are derived from only one genre sub-corpus. 6. 'New sense' is treated as one of the sense classes.

- JAIST-3
  The system is an ensemble of JAIST-1 and JAIST-2. The judgment of a new sense is performed by JAIST-1. The output of JAIST-1 is chosen when the similarity between a cluster and a sense in the dictionary is sufficiently high. Otherwise, the output of JAIST-2 is used.

- MSS-1,2,3
  1. Maximum entropy, 2. Three word forms/lemmas/POSs before or after the target word, bigrams, and skip bigrams in the context, bag-of-words in the document, a class of the document categorized by a topic classifier, etc. 3. None, 4. None, 5. For each target word, the system selected the genre and dictionary examples combinations for training data, which got the best results in cross-validation. 6. The system calculated the entropy for each target word given by the Maximum Entropy Model (MEM). It assumed that high entropy (when probabilities of classes are uniformly dispersed) was indicative of a new sense. The threshold was tuned by using the words with a new sense tag in the training data. Three official submissions correspond to different thresholds.

- RALI-1, RALI-2
  1. Naive Bayes, 2. Only the 'writing' of the words (inside of <mor> tag), 3. The Mainichi 2005 corpus of NTCIR, parsed with chasen+unidic, 4. None, 5. Not taken into account, 6. 'New sense' is only used when it is evident in the training data

For more details, please refer to their description papers.

6 Their Results

The evaluation results of all the systems are shown in tables 1 and 2. "Baseline" for WSD indicates the results of the baseline system that used SVM with the following features:

- Morphological features
  Bag-of-words (BOW), Part-of-speech (POS), and detailed POS classification. We extract these features from the target word itself and the two words to the right and left of it.

- Syntactic features
  - If the POS of a target word is a noun, extract the verb in a grammatical dependency relation with the noun.

Table 1: Results: Word sense disambiguation

	Precision
Baseline	0.7528
HIT-1	0.6612
JAIST-1	0.6864
JAIST-2	0.7476
JAIST-3	0.7208
MSS-1	0.6404
MSS-2	0.6384
MSS-3	0.6604
RALI-1	0.7592
RALI-2	0.7636

Table 2: Results: New sense detection

	Accuracy	Precision	Recall
Baseline	0.9844	-	0
HIT-1	0.9132	0.0297	0.0769
JAIST-1	0.9512	0.0337	0.0769
JAIST-2	0.9872	1	0.1795
JAIST-3	0.9532	0.0851	0.2051
MSS-1	0.9416	0.1409	0.5385
MSS-2	0.9384	0.1338	0.5385
MSS-3	0.9652	0.2333	0.5385
RALI-1	0.9864	0.7778	0.1795
RALI-2	0.9872	0.8182	0.2308

  - If the POS of a target word is a verb, extract the noun in a grammatical dependency relation with the verb.

- Figures in Bunrui-Goi-Hyou
  4 and 5 digits regarding the content word to the right and left of the target word.

The baseline system did not take into account any information on the text genre. "Baseline" for new sense detection (NSD) indicates the results of the baseline system, which outputs a sense in the dictionary and never outputs the new sense tag. Precision and recall for NSD are shown just for reference. Because relatively few instances for a new word sense were found (39 out of 2500), the task of the new sense detection was found to be rather difficult.

Tables 3 and 4 show the results for nouns, verbs, and adjectives. In our comparison of the baseline system scores for WSD, the score for nouns was the biggest, and the score for verbs was the smallest (table 3). However, the average entropy of nouns was the second biggest (0.7257), and that of verbs was the biggest (1.194)[5].

Table 3: Results for each POS (Precision): Word sense disambiguation

	Noun	Verb	Adjective
Baseline	0.8255	0.6878	0.732
HIT-1	0.7436	0.5739	0.7
JAIST-1	0.7645	0.5957	0.76
JAIST-2	0.84	0.6626	0.732
JAIST-3	0.8236	0.6217	0.724
MSS-1	0.7	0.5504	0.792
MSS-2	0.6991	0.5470	0.792
MSS-3	0.7218	0.5713	0.8
RALI-1	0.8236	0.6965	0.764
RALI-2	0.8127	0.7191	0.752

Table 4: Results for each POS (Accuracy): New sense detection

	Noun	Verb	Adjective
Baseline	0.97	0.9948	1
HIT-1	0.8881	0.9304	0.944
JAIST-1	0.9518	0.9470	0.968
JAIST-2	0.9764	0.9948	1
JAIST-3	0.9564	0.9470	0.968
MSS-1	0.9355	0.9409	0.972
MSS-2	0.9336	0.9357	0.972
MSS-3	0.96	0.9670	0.98
RALI-1	0.9745	0.9948	1
RALI-2	0.9764	0.9948	1

We set up three word classes, $D_{diff}(E(w) \geq 1)$, $D_{mid}(0.5 \leq E(w) < 1)$, and $D_{easy}(E(w) < 0.5)$. D_{diff}, D_{mid}, and D_{easy} consist of 20, 19 and 11 words, respectively. Tables 5 and 6 show the results for each word class. The results of WSD are quite natural in that the higher $E(w)$ is, the more difficult WSD is, and the more the performance degrades.

7 Conclusion

This paper reported an overview of the SemEval-2 Japanese WSD task. The data used in this task will be available when you contact the task organizer and sign a copyright agreement form. We hope this valuable data helps many researchers improve their WSD systems.

[5]The average entropy of adjectives was 0.6326.

Table 5: Results for entropy classes (Precision): Word sense disambiguation

	D_{easy}	D_{mid}	D_{diff}
Baseline	0.9418	0.7411	0.66
HIT-1	0.8436	0.6832	0.54
JAIST-1	0.8782	0.7158	0.553
JAIST-2	0.9509	0.7484	0.635
JAIST-3	0.92	0.7368	0.596
MSS-1	0.8291	0.6558	0.522
MSS-2	0.8273	0.6558	0.518
MSS-3	0.8345	0.6905	0.536
RALI-1	0.9455	0.7653	0.651
RALI-2	0.94	0.7558	0.674

Table 6: Results for Entropy classes (Accuracy): New sense detection

	D_{easy}	D_{mid}	D_{diff}
Baseline	1	0.9737	0.986
HIT-1	0.8909	0.9095	0.929
JAIST-1	0.9672	0.9505	0.943
JAIST-2	1	0.9811	0.986
JAIST-3	0.9673	0.9558	0.943
MSS-1	0.9818	0.9221	0.938
MSS-2	0.98	0.9221	0.931
MSS-3	0.9873	0.9611	0.957
RALI-1	1	0.9789	0.986
RALI-2	1	0.9811	0.986

Acknowledgments

We would like to thank all the participants and the annotators for constructing this sense tagged corpus.

References

Eneko Agirre and Oier Lopez de Lacalle. 2008. On robustness and domain adaptation using svd for word sense disambiguation. In *Proc. of COLING'08*.

Yee Seng Chang and Hwee Tou Ng. 2006. Estimating class priors in domain adaptation for wsd. In *Proc. of ACL'06*.

Ido Dagan, Lillian Lee, and Fernando Pereira. 1997. Similarity-based methods for word sense disambiguation. In *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 56–63.

A. Kilgarriff and J. Rosenzweig. 2000. English senseval: Report and results. In *Proc. LREC'00*.

J. Lin. 1991. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151.

Kikuo Maekawa. 2008. Balanced corpus of contemporary written japanese. In *Proceedings of the 6th Workshop on Asian Language Resources (ALR)*, pages 101–102.

National Institute of Japanese Language. 1964. *Bunruigoihyou*. Shuuei Shuppan. In Japanese.

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

Kiyoaki Shirai. 2001. Senseval-2 japanese dictionary task. In *Proceedings of SENSEVAL-2: Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 33–36.