An overview of the SemEval-2 Japanese WSD task is presented. The new characteristics of our task are (1) the task will use the first balanced Japanese sense-tagged corpus, and (2) 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. 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.

**Key Words:** Word Sense Disambiguation, Balanced Corpus, Text Genre, New Word Sense

### 1 Introduction

This paper reports an overview of the SemEval-2 Japanese Word Sense Disambiguation (WSD) task. It is an extension of the SENSEVAL-2 Japanese monolingual dictionary-based task (Shirai 2001), and is a lexical sample task. In the usual WSD task, participants received a dictionary as a sense inventory and training data, and they are requested to annotate senses from the dictionary for target words in test data.

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 the Brown corpus and the BNC corpus. The first balanced corpus of contemporary written Japanese (BCCWJ corpus) is now being constructed as part of a government funded 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 data. 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.

Word senses are defined according to the Iwanami Kokugo Jiten (Nishio, Iwabuchi, and Mizutani 1994), a Japanese dictionary published by Iwanami Shoten. The training data consists of three genres (books, newspaper articles, and white papers) and is manually annotated with sense IDs. 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\(^2\)) with marked target words as the 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, consisting 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 explains two evaluation methodologies in our task. Section five briefly introduces participating systems, and section six describes their results. Finally, section seven concludes the paper.

2 Data

For the task, we prepared three types of data: a sense inventory, training data, and test data. In the following subsections, we describe detailed statistics and specifications of the sense inventory and the data.

\(^2\) Yahoo!知恵袋 (http://chiebukuro.yahoo.co.jp/)
2.1 Sense inventory

As described in section one, word senses are defined according to a Japanese dictionary, the Iwanami Kokugo Jiten. The total number of headwords and the word senses in the Iwanami Kokugo Jiten is 60,321 and 85,870, respectively. For ambiguous words that have multiple word senses, the number of headwords and the average number of word senses for them is 12,360 and 3.067, respectively.

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 levels. Figure 1 shows a sample entry of the Iwanami Kokugo Jiten.

2.2 BCCWJ corpus

BCCWJ corpus is the first balanced corpus of contemporary written Japanese, that is now being constructed. We are now constructing a sense-tagged corpus based on it. In this subsection, we briefly outline BCCWJ corpus before we explain our sense-tagged corpus.

The volume of the corpus will be 100 million words in total. The BCCWJ consists of the following 3 sub-corpora (Maekawa 2008):

(1) Publication sub-corpus

This sub-corpus represents the production, as opposed to the reception aspect of contem-
porary Japanese. It consists of 35 million words from books, magazines, and newspapers.

(2) Library sub-corpus

The sampling for this sub-corpus was from the books in at least 13 public libraries in Tokyo. It consists of 30 million words.

(3) Special-purpose sub-corpus

This corpus is the aggregate of various mini corpora. The mini corpora include texts of governmental white papers, Web texts, and so on. The whole corpus consists of 35 million words.

The subset of the BCCWJ corpus was annotated with morphological information for all words, as explained in the next subsection, and the subset data was named ‘core data’. We manually annotated word sense IDs for all ambiguous words in the core data. The core data consists of 200 thousand words from books, 200 thousand words from white papers, 300 thousand words from newspaper articles, and 100 thousand words from a Q&A site on the WWW.

2.3 Annotation for the data

The annotated information both for the training and test 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 a morphological analyzer ChaSen\(^3\) with a short-unit dictionary named UniDic\(^4\) and was manually post-edited.

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

- Word sense IDs
  Word sense IDs were manually annotated\(^5\). The details are described in section three.

2.4 Training and test data

The training data consists of 240 documents of three genres (books, newspaper articles, and white papers) from the BCCWJ corpus. 3,437 word tokens 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:

\(^3\) http://chasen-legacy.sourceforge.jp/
\(^4\) http://www.tokuteicorpus.jp/dist/
\(^5\) They were hidden from the participants in the test data during the formal run.
(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.

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.

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.

2.5 Statistics of training/test data

First, we listed 50 target words in the appendix. In the list, we also showed their POSs and their word classes by the entropy of the word sense distribution, that we will mention in section seven. ‘N’, ‘v’, and ‘a’ are an abbreviation of noun, verb, and adjective, respectively.

The number of instances for a new word sense in the training and test data is 15 and 38, respectively.

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 and Linguistics. 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 set 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

---

6 The word sense IDs for them were hidden from the participants.
7 Due to space limits, we unfortunately cannot present the Jensen Shannon (JS) divergence (Lin 1991; Dagan, Lee, and Pereira 1997) between the word sense distributions of two different genres.
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 word sense IDs for the training and test data. All of them majored in linguistics and 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 them carefully and judged whether those instances really had a new sense.

The inter-annotator agreement between the two annotators in step (1) was calculated with Kappa statistics for a fragment of the corpus tagged by multiple annotators in a preliminary annotation, and it was 0.678.

4 Evaluation methodology

The evaluation was returned in the following two ways:

(1) The sense IDs outputted by a system 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 on the basis of the following six aspects:

1. learning algorithm used,
2. features used,
3. language resources used,
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,
  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-1,2**
  (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 outputted only for the words that have instances for a new word sense in the training data.

For more details, please refer to their description papers in the SemEval-2010 workshop (Brosseau-Villeneuve, Kando, and Nie 2010; Shirai and Nakamura 2010; Fujita, Duh, Fujino, Taira, and Shindo 2010).

### 6 Results

The evaluation results of all the systems are shown in Tables 1 and 2. “MFS” for WSD indicates the results of the baseline that outputed the ‘most frequent sense’ in the training data. “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.
Table 1  Results (Precision): Word sense disambiguation.

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFS</td>
<td>0.6896</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.7528</td>
</tr>
<tr>
<td>HIT-1</td>
<td>0.6612</td>
</tr>
<tr>
<td>JAIST-1</td>
<td>0.6864</td>
</tr>
<tr>
<td>JAIST-2</td>
<td>0.7476</td>
</tr>
<tr>
<td>JAIST-3</td>
<td>0.7208</td>
</tr>
<tr>
<td>MSS-1</td>
<td>0.6404</td>
</tr>
<tr>
<td>MSS-2</td>
<td>0.6384</td>
</tr>
<tr>
<td>MSS-3</td>
<td>0.6604</td>
</tr>
<tr>
<td>RALI-1</td>
<td>0.7592</td>
</tr>
<tr>
<td>RALI-2</td>
<td>0.7636</td>
</tr>
</tbody>
</table>

Table 2  Results (Accuracy): New sense detection.

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.9844</td>
<td>—</td>
<td>0</td>
</tr>
<tr>
<td>HIT-1</td>
<td>0.9132</td>
<td>0.0297</td>
<td>0.0769</td>
</tr>
<tr>
<td>JAIST-1</td>
<td>0.9512</td>
<td>0.0337</td>
<td>0.0769</td>
</tr>
<tr>
<td>JAIST-2</td>
<td>0.9872</td>
<td>1</td>
<td>0.1795</td>
</tr>
<tr>
<td>JAIST-3</td>
<td>0.9532</td>
<td>0.0851</td>
<td>0.2051</td>
</tr>
<tr>
<td>MSS-1</td>
<td>0.9416</td>
<td>0.1409</td>
<td>0.5385</td>
</tr>
<tr>
<td>MSS-2</td>
<td>0.9384</td>
<td>0.1338</td>
<td>0.5385</td>
</tr>
<tr>
<td>MSS-3</td>
<td>0.9652</td>
<td>0.2333</td>
<td>0.5385</td>
</tr>
<tr>
<td>RALI-1</td>
<td>0.9864</td>
<td>0.7778</td>
<td>0.1795</td>
</tr>
<tr>
<td>RALI-2</td>
<td>0.9872</td>
<td>0.8182</td>
<td>0.2308</td>
</tr>
</tbody>
</table>

- **Syntactic features**
  - If the POS of a target word is a noun, extract the verb in a grammatical dependency relation with the noun.
  - 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 (38 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)\(^8\).

We set up three word classes, \(D_{diff}(E(w) \geq 1)\), \(D_{mid}(0.5 \leq E(w) < 1)\), and \(D_{easy}(E(w) < 0.5)\), based on the relative “difficulty” in disambiguating the sense of words that we mentioned in section 3.1. \(D_{diff}\), \(D_{mid}\), and \(D_{easy}\) consist of 20, 19 and 11 words, respectively. Tables 5 and

\(^8\) The average entropy of adjectives was 0.6326.
Table 3  Results for each POS (Precision):
Word sense disambiguation.

<table>
<thead>
<tr>
<th></th>
<th>Noun</th>
<th>Verb</th>
<th>Adjective</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFS</td>
<td>0.7773</td>
<td>0.6043</td>
<td>0.696</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.8255</td>
<td>0.6878</td>
<td>0.732</td>
</tr>
<tr>
<td>HIT-1</td>
<td>0.7436</td>
<td>0.5739</td>
<td>0.7</td>
</tr>
<tr>
<td>JAIST-1</td>
<td>0.7645</td>
<td>0.5957</td>
<td>0.76</td>
</tr>
<tr>
<td>JAIST-2</td>
<td>0.84</td>
<td>0.6626</td>
<td>0.732</td>
</tr>
<tr>
<td>JAIST-3</td>
<td>0.8236</td>
<td>0.6217</td>
<td>0.724</td>
</tr>
<tr>
<td>MSS-1</td>
<td>0.7</td>
<td>0.5504</td>
<td>0.792</td>
</tr>
<tr>
<td>MSS-2</td>
<td>0.6991</td>
<td>0.5470</td>
<td>0.792</td>
</tr>
<tr>
<td>MSS-3</td>
<td>0.7218</td>
<td>0.5713</td>
<td>0.8</td>
</tr>
<tr>
<td>RALI-1</td>
<td>0.8236</td>
<td>0.6965</td>
<td>0.764</td>
</tr>
<tr>
<td>RALI-2</td>
<td>0.8127</td>
<td>0.7191</td>
<td>0.752</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>Noun</th>
<th>Verb</th>
<th>Adjective</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.97</td>
<td>0.9948</td>
<td>1</td>
</tr>
<tr>
<td>HIT-1</td>
<td>0.8881</td>
<td>0.9304</td>
<td>0.944</td>
</tr>
<tr>
<td>JAIST-1</td>
<td>0.9518</td>
<td>0.9470</td>
<td>0.968</td>
</tr>
<tr>
<td>JAIST-2</td>
<td>0.9764</td>
<td>0.9948</td>
<td>1</td>
</tr>
<tr>
<td>JAIST-3</td>
<td>0.9564</td>
<td>0.9470</td>
<td>0.968</td>
</tr>
<tr>
<td>MSS-1</td>
<td>0.9355</td>
<td>0.9409</td>
<td>0.972</td>
</tr>
<tr>
<td>MSS-2</td>
<td>0.9336</td>
<td>0.9357</td>
<td>0.972</td>
</tr>
<tr>
<td>MSS-3</td>
<td>0.96</td>
<td>0.9670</td>
<td>0.98</td>
</tr>
<tr>
<td>RALI-1</td>
<td>0.9745</td>
<td>0.9948</td>
<td>1</td>
</tr>
<tr>
<td>RALI-2</td>
<td>0.9764</td>
<td>0.9948</td>
<td>1</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>$D_{easy}$</th>
<th>$D_{mid}$</th>
<th>$D_{diff}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFS</td>
<td>0.9509</td>
<td>0.6821</td>
<td>0.553</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.9418</td>
<td>0.7411</td>
<td>0.66</td>
</tr>
<tr>
<td>HIT-1</td>
<td>0.8436</td>
<td>0.6832</td>
<td>0.54</td>
</tr>
<tr>
<td>JAIST-1</td>
<td>0.8782</td>
<td>0.7158</td>
<td>0.553</td>
</tr>
<tr>
<td>JAIST-2</td>
<td>0.9509</td>
<td>0.7484</td>
<td>0.635</td>
</tr>
<tr>
<td>JAIST-3</td>
<td>0.92</td>
<td>0.7368</td>
<td>0.596</td>
</tr>
<tr>
<td>MSS-1</td>
<td>0.8291</td>
<td>0.6558</td>
<td>0.522</td>
</tr>
<tr>
<td>MSS-2</td>
<td>0.8273</td>
<td>0.6558</td>
<td>0.518</td>
</tr>
<tr>
<td>MSS-3</td>
<td>0.8345</td>
<td>0.6905</td>
<td>0.536</td>
</tr>
<tr>
<td>RALI-1</td>
<td>0.9455</td>
<td>0.7653</td>
<td>0.651</td>
</tr>
<tr>
<td>RALI-2</td>
<td>0.94</td>
<td>0.7558</td>
<td>0.674</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>$D_{easy}$</th>
<th>$D_{mid}$</th>
<th>$D_{diff}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>1</td>
<td>0.9737</td>
<td>0.986</td>
</tr>
<tr>
<td>HIT-1</td>
<td>0.8909</td>
<td>0.9095</td>
<td>0.929</td>
</tr>
<tr>
<td>JAIST-1</td>
<td>0.9672</td>
<td>0.9505</td>
<td>0.943</td>
</tr>
<tr>
<td>JAIST-2</td>
<td>1</td>
<td>0.9811</td>
<td>0.986</td>
</tr>
<tr>
<td>JAIST-3</td>
<td>0.9673</td>
<td>0.9558</td>
<td>0.943</td>
</tr>
<tr>
<td>MSS-1</td>
<td>0.9818</td>
<td>0.9221</td>
<td>0.938</td>
</tr>
<tr>
<td>MSS-2</td>
<td>0.98</td>
<td>0.9221</td>
<td>0.931</td>
</tr>
<tr>
<td>MSS-3</td>
<td>0.9873</td>
<td>0.9611</td>
<td>0.957</td>
</tr>
<tr>
<td>RALI-1</td>
<td>1</td>
<td>0.9789</td>
<td>0.986</td>
</tr>
<tr>
<td>RALI-2</td>
<td>1</td>
<td>0.9811</td>
<td>0.986</td>
</tr>
</tbody>
</table>

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.

Lastly, Tables 7 and 8 show the results for each genre of instances. Test data consists of four genres of instances: books (PB), newspaper articles (PN), white papers (OW), and documents from a Q&A site on the WWW (OC).

The results of WSD are quite natural in that the performance for OC sharply degrades compared with other genres, since training data does not contain any instances from the documents from a Q&A site on the WWW (OC).
Table 7 Results for each genre (Precision): Word sense disambiguation.

<table>
<thead>
<tr>
<th></th>
<th>OC</th>
<th>OW</th>
<th>PB</th>
<th>PN</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFS</td>
<td>0.5587</td>
<td>0.8937</td>
<td>0.7104</td>
<td>0.7539</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.6413</td>
<td>0.9324</td>
<td>0.7738</td>
<td>0.8028</td>
</tr>
<tr>
<td>HIT-1</td>
<td>0.5742</td>
<td>0.7198</td>
<td>0.6821</td>
<td>0.7192</td>
</tr>
<tr>
<td>JAIST-1</td>
<td>0.6232</td>
<td>0.8261</td>
<td>0.6719</td>
<td>0.7382</td>
</tr>
<tr>
<td>JAIST-2</td>
<td>0.6439</td>
<td>0.9082</td>
<td>0.7692</td>
<td>0.7918</td>
</tr>
<tr>
<td>JAIST-3</td>
<td>0.5948</td>
<td>0.9082</td>
<td>0.7421</td>
<td>0.7839</td>
</tr>
<tr>
<td>MSS-1</td>
<td>0.6142</td>
<td>0.8164</td>
<td>0.6097</td>
<td>0.6577</td>
</tr>
<tr>
<td>MSS-2</td>
<td>0.6103</td>
<td>0.8164</td>
<td>0.6086</td>
<td>0.6562</td>
</tr>
<tr>
<td>MSS-3</td>
<td>0.6426</td>
<td>0.8647</td>
<td>0.6199</td>
<td>0.6719</td>
</tr>
<tr>
<td>RALI-1</td>
<td>0.6968</td>
<td>0.9179</td>
<td>0.7545</td>
<td>0.7902</td>
</tr>
<tr>
<td>RALI-2</td>
<td>0.7032</td>
<td>0.8937</td>
<td>0.7613</td>
<td>0.7981</td>
</tr>
</tbody>
</table>

Table 8 Results for each genre (Accuracy): New sense detection.

<table>
<thead>
<tr>
<th></th>
<th>OC</th>
<th>OW</th>
<th>PB</th>
<th>PN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.9987</td>
<td>0.9952</td>
<td>0.9898</td>
<td>0.9558</td>
</tr>
<tr>
<td>HIT-1</td>
<td>0.9471</td>
<td>0.8213</td>
<td>0.9152</td>
<td>0.8991</td>
</tr>
<tr>
<td>JAIST-1</td>
<td>0.9148</td>
<td>0.9952</td>
<td>0.9762</td>
<td>0.9464</td>
</tr>
<tr>
<td>JAIST-2</td>
<td>0.9987</td>
<td>0.9952</td>
<td>0.9932</td>
<td>0.9621</td>
</tr>
<tr>
<td>JAIST-3</td>
<td>0.9148</td>
<td>0.9952</td>
<td>0.9785</td>
<td>0.9511</td>
</tr>
<tr>
<td>MSS-1</td>
<td>0.9548</td>
<td>0.9420</td>
<td>0.9434</td>
<td>0.9227</td>
</tr>
<tr>
<td>MSS-2</td>
<td>0.9497</td>
<td>0.9420</td>
<td>0.9412</td>
<td>0.9196</td>
</tr>
<tr>
<td>MSS-3</td>
<td>0.9897</td>
<td>0.9903</td>
<td>0.9548</td>
<td>0.9416</td>
</tr>
<tr>
<td>RALI-1</td>
<td>0.9987</td>
<td>0.9952</td>
<td>0.9932</td>
<td>0.9590</td>
</tr>
<tr>
<td>RALI-2</td>
<td>0.9987</td>
<td>0.9952</td>
<td>0.9921</td>
<td>0.9637</td>
</tr>
</tbody>
</table>

7 Conclusion

This paper reported an overview of the SemEval-2 Japanese WSD task. 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. 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.

Appendix

(1) 「相手」 (n, \(D_{mid}\)).
(2) 「合う」 (v, \(D_{diff}\)).
(3) 「上げる」 (v, \(D_{diff}\)).
(4) 「与える」 (v, \(D_{diff}\)).
(5) 「生きる」 (v, \(D_{easy}\)).
(6) 「意味」 (n, \(D_{diff}\)).
(7) 「入れる」 (v, \(D_{diff}\)).
(8) 「大きい」 (a, \(D_{easy}\)).
(9) 「教える」 (v, \(D_{mid}\)).
(10) 「可能」 (n, \(D_{mid}\)).
(11) 「考える」(v, $D_{easy}$).
(12) 「関係」(n, $D_{mid}$).
(13) 「技術」(n, $D_{mid}$).
(14) 「経済」(n, $D_{easy}$).
(15) 「現場」(n, $D_{mid}$).
(16) 「子供」(n, $D_{mid}$).
(17) 「時間」(n, $D_{mid}$).
(18) 「市場」(n, $D_{diff}$).
(19) 「社会」(n, $D_{mid}$).
(20) 「情報」(n, $D_{mid}$).
(21) 「進める」(v, $D_{diff}$).
(22) 「する」(v, $D_{diff}$).
(23) 「高い」(a, $D_{mid}$).
(24) 「出す」(v, $D_{diff}$).
(25) 「立つ」(v, $D_{diff}$).
(26) 「強い」(a, $D_{easy}$).
(27) 「手」(n, $D_{diff}$).
(28) 「出る」(v, $D_{diff}$).
(29) 「電話」(n, $D_{mid}$).
(30) 「取る」(v, $D_{diff}$).
(31) 「乗る」(v, $D_{diff}$).
(32) 「場合」(n, $D_{mid}$).
(33) 「入る」(v, $D_{diff}$).
(34) 「初め」(n, $D_{mid}$).
(35) 「始める」(v, $D_{mid}$).
(36) 「場所」(n, $D_{easy}$).
(37) 「早い」(a, $D_{mid}$).
(38) 「一つ」(n, $D_{easy}$).
(39) 「開く」(v, $D_{easy}$).
(40) 「文化」(n, $D_{easy}$).
(41) 「外」(n, $D_{easy}$).
(42) 「前」(n, $D_{diff}$).
(43) 「見える」(v, $D_{diff}$).
(44) 「認める」(v, $D_{diff}$).
(45) 「見る」 (v, $D_{diff}$).
(46) 「持つ」 (v, $D_{diff}$).
(47) 「求める」 (v, $D_{mid}$).
(48) 「もの」 (n, $D_{mid}$).
(49) 「やる」 (v, $D_{easy}$).
(50) 「良い」 (a, $D_{mid}$)

Acknowledgement

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

Reference

Publisher. In Japanese.
2: Second International Workshop on Evaluating Word Sense Disambiguation Systems,
pp. 33–36.
Shirai, K. and Nakamura, M. (2010). “JAIST: Clustering and Classification Based Approaches for
Japanese WSD.” In Proceedings of the 5th International Workshop on Semantic Evaluation,

奥村 学：1962年生。1984年東京工業大学工学部情報工学科卒業。1989年
同大学院博士課程修了。同年。東京工業大学工学部情報工学科助手。1992年
北陸先端科学技术大学院大学情報科学研究科助教授。2000年東京工業大学精密
工学研究所助教授。2009年同教授。現在に至る。工学博士。自然言語処理,
知的情報提示技術。語学学習支援。テキスト評価分析。テキストマイニングに
に関する研究に従事。情報処理学会。電子情報通信学会。人工知能学会。AAAI。言語
処理学会。ACL。認知科学会。計量国語学会各会員。
白井 清昭：1993年東京工業大学工学部情報工学科卒業。1998年同大学院情報
理工学研究科博士後期課程修了。同年同大学院情報理工学研究科計算工学専
攻助手。2001年北陸先端科学技术大学院大学情報科学研究科助教授。現在同
准教授。博士（工学）。自然言語処理に関する研究に従事。言語処理学会。情
報処理学会。人工知能学会。ACL各会員。
古宮信那子：2005年東京農工大学工学部情報コミュニケーション工学科卒。2009
年同大学院博士後期課程電子情報工学専攻修了。博士（工学）。同年東京工
業大学精密工学研究所研究員。2010年東京農工大学工学研究院特任助教。現
在に至る。自然言語処理の研究に従事。情報処理学会。人工知能学会。言語
処理学会各会員。
横野 克：2003年岡山大学工学部情報工学科卒。2008年同大学院自然科学研究
科産業創成工学専攻単位取得退学。同年東京工業大学精密工学研究所研
究員。現在に至る。博士（工学）。自然言語処理の研究に従事。情報処理学会。
言語処理学会各会員.