Study of Word Sense Disambiguation System that uses Contextual Features
- Approach of Combining Associative Concept Dictionary and Corpus -

Kyota Tsutsumida*, Jun Okamoto†, Shun Ishizaki‡,
Makoto Nakatsuji*, Akimichi Tanaka*, Tadasu Uchiyama*

*NTT Cyber Solutions Laboratories, NTT Corporation
†Keio Research Institute at SFC
‡Graduate School of Media and Governance, Keio University
*1-1 Hikari-no-oka Yokosuka-Shi Kanagawa, 239-0847, Japan
† ‡5322 Endo Fujisawa-Shi Kanagawa, 252-8520, Japan
*{tsutsumida.kyota, nakatsuji.makoto, tanaka.akimichi, uchiyama.tadasu}@lab.ntt.co.jp,
{†juno, ‡ishizaki}@sfc.keio.ac.jp

Abstract

We propose a Word Sense Disambiguation (WSD) method that accurately classifies ambiguous words to concepts in the Associative Concept Dictionary (ACD) even when the test corpus and the training corpus for WSD are acquired from different domains. Many WSD studies determine the context of the target ambiguous word by analyzing sentences containing the target word. However, they offer poor performance when they are applied to a corpus that differs from the training corpus. One solution is to use associated words that are domain-independently assigned to the concept in ACD; i.e. many users commonly imagine those words against a given concept. Furthermore, by using the associated words of a concept as search queries for a training corpus, our method extracts relevant words, those that are computationally judged to be related to that concept. By checking the frequency of associated words and relevant words that appear near to the target word in a sentence in the test corpus, our method classifies the target word to the concept in ACD. Our evaluation using two different types of corpus demonstrates its good accuracy.

1. Introduction

In many natural languages, a word can have multiple meanings and an effective Word Sense Disambiguation (WSD) system is needed to determine the intended meaning in the given sentence or context. WSD systems contribute to the performance improvement of many Natural Language Processing (NLP) systems, such as machine translation systems and information retrieval systems.

In order to disambiguate word sense, past studies have taken the machine learning approach using contextual features against a target ambiguous word in a particular sentence. The contextual features against a target ambiguous word include co-occurrence words with the target word or word dependency relationship against the target word in the sentence. In the Japanese Dictionary Task in SENSE-VAL2 (Kurohashi and Shirai, 2001), the best system (Murata et al., 2003) uses various contextual features such as co-occurrence words extracted by using the Japanese morphological analyzer, JUMAN[1], and the word dependency relationship analyzed by using the parser, KNP[2]. Shirai and Yagi (2004) showed that by using hypernyms in definition sentences of the EDR concept dictionary (1995) in addition to the contextual features used by (Murata et al., 2003), their method offered better performance than Murata's. However, there may be not much difference between these methods, especially in terms of applicability (portability). Okamoto and Ishizaki (2007) proposed a WSD method that uses the Contextual Dynamic Network Model with the Associative Concept Dictionary (Okamoto and Ishizaki, 2001) which includes ontological information; each concept has several associated words, which are domain-independently assigned to the concept in ACD only when many users commonly imagine those words against the concept.

As we know, WSD systems need a lot of data for training, otherwise their accuracy is poor. Our approach is to automatically extract effective relevant words, those that are computationally judged to be related to the concept in ACD, and use those relevant words as one of the contextual features against the target words. Moreover, we integrate the contextual features, which are almost the same as those in Murata's method (Murata et al., 2003) and extracted relevant words from the corpus as extended contextual feature sets.

The proposed WSD method is used in combination with a training corpus and the Associative Concept Dictionary. By using the ACD entries, we can extract a lot of relevant words from the corpus. By increasing the number of relevant words for each concept in ACD, the WSD system can deal with contextual features more appropriately, and the accuracy of the WSD system is expected to be improved. In Section 2, we describe ACD, two types of corpora, target words, and labels. Section 3 describes the proposed method. Section 4 describes the methods, Baseline and our method, and the experiments conducted on them. Section 5 introduces the results of evaluation experiments and analyzes our method. Our method is shown to offer better versatility, applicability (portability) and higher effectiveness than the existing alternatives.

---

[1]http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman-e.html

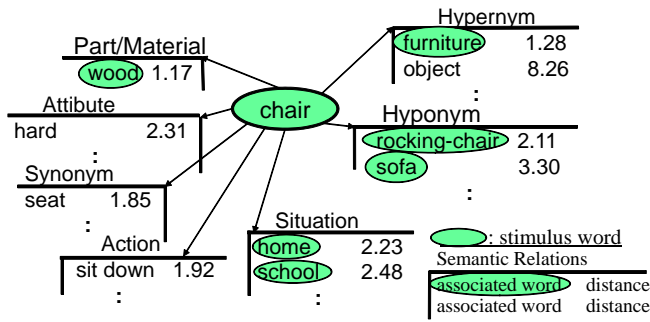[2]http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp-e.html

Figure 1: Conceptual diagram of ACD. "chair" is the stimulus word for the experiment. Some associated words are also indexed as stimulus word in ACD.
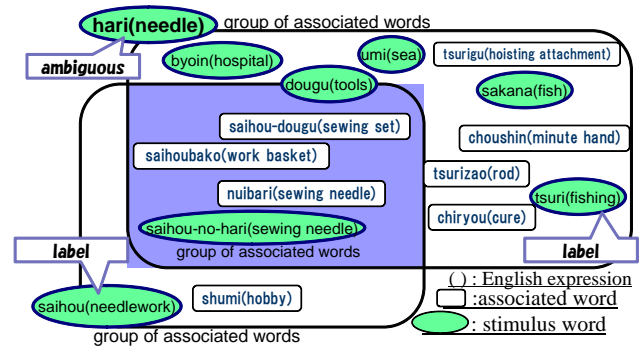

Figure 2: The group of the associated words in ACD and ACD-for-WSD. The shaded region is the associated words of "saihou-no-hari(sewing needle)" which is the stimulus used in ACD-for-WSD.

2. Associative Concept Dictionary and Corpora

2.1. Associative Concept Dictionary (ACD)

The Associative Concept Dictionary (Okamoto and Ishizaki, 2001) has been built based on the results of large-scale online association experiments. It first has stimulus words, and then creates concepts collaboratively by collecting associated words imagined by a lot of users, thus a concept in ACD are defined as a set of associated words that are imagined by presenting stimulus words to a lot of users.

It calculates the association distances between the stimulus words and the associated words from the frequency, order of association words, and the time delay until the association is made. ACD is organized in a hierarchical structure in terms of hypernyms and hyponyms. Attribute information is used to explain the features of any given word. In an association experiment, each stimulus word was presented to 50 subjects, students of Keio University. The number of stimulus words is currently 1,100. Total number of associated words with overlap is about 280,000, and the number of associated words free of overlap is about 64,000.

In our method, we try to classify the sentence that includes the target ambiguous word to concepts in ACD. In other words, we analyze the context of the target ambiguous word in the sentence according to the concept in ACD.

2.2. Additional Experiment on ACD-for-WSD

Our method uses stimulus words and associated words to extract relevant words (extraction method is described in Section 3). If we only submit stimulus words as queries for the training corpus to extract relevant words, the search results are likely to include many irrelevant documents. One approach is to collect associated words by submitting both stimulus word and one of the associated words.

The stimulus words in ACD have multiple contexts. In the case of "hari(needle)", the group of associated words of "hari(needle)" covers several contexts such as "saihou(needlework)" and "tsuri(fishing)" as shown in Figure 2. To disambiguate the stimulus word in the process of the word association experiment, we also built the ACD for WSD by submitting a phrase that has both the ambiguous word (stimulus word in ACD) and context (class label

word) at the same time such as "saihou-no-hari(sewing needle)". The shaded region in Figure 2, the associated words from "saihou-no-hari(sewing needle)", represents the intersection of ambiguous word ("needle") and the class label ("needlework").

In this additional experiment, each stimulus phrase had 20 subjects, students of Keio University.

2.3. Corpus (Mainichi Newspaper / Aozora Bunko[3])

We prepared two different corpora in order to evaluate both the versatility and effectiveness of our method. The Mainichi newspaper corpus (1993 to 1995) was used to train and test the system. The sentences including the target word (ambiguous word) were manually tagged with the correct concept labels (proper meaning in the context). We also extracted and tagged the sentences of the Aozora Bunko corpus, a collection of old Japanese novels.

Aozora Bunko corpus is used only to test the system, not for training the system.

In the evaluation process, the WSD systems were trained on the Mainichi newspaper corpus but evaluated on both the Mainichi newspaper corpus and the Aozora bunko corpus. In our evaluation using Mainichi newspaper's dataset, the accuracy was achieved by 10-fold cross validation. On the other hand, the Aozora bunko corpus was evaluated by the classifier trained on the Mainichi newspaper corpus.

3. Proposed Method (Relevant Word Extraction in Combination with Corpora and ACD)

To provide the WSD system with common knowledge about words and contexts, we used the associated words in ACD or ACD-for-WSD, those words that have an associative relationship with each context into which each polysemous word should be classified.

Second, by using the stimulus words (label words of target words) and associated words as search queries, the system extracted a lot of sentences containing the associated words. In this paper, we selected the sentences with the stimulus words and 5 or more associated words for each
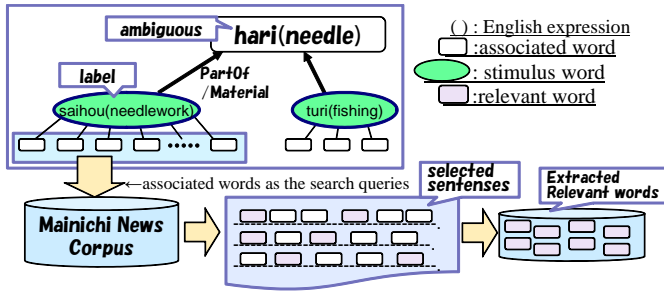
---

[3]http://www.aozora.gr.jp/

Figure 3: Procedure of extracting relevant words in the case of ACD.

label (context). These selected sentences were analyzed by the Japanese morphological analyzer JUMAN, and the words in the sentences were taken to be relevant words in terms of context.

The above approach allows the system to collect a lot of relevant words about each context from the corpus. The proposed system uses these relevant words as the expanded contextual feature sets and integrates other contextual features that are acquired by analyzing the input sentence with the target word. In the case of ACD-for-WSD, the WSD system uses the same sequences to collect relevant words and expand the contextual feature sets.

4. Methods and Experiments

In this paper, we try to classify input sentences, which include the target words (ambiguous words) into the proper class labels. In the evaluation process, we compared the accuracies of five methods including the 2 proposed methods. Each method took the same machine learning approach (Naive Bayes method) but used different contextual feature sets.

4.1. Methods

Baseline (most freq sense)

> This method always selects the most frequently used sense. This is often used as the Baseline in WSD studies (Murata et al., 2003).

CRL (Murata, 2003)

> This WSD classifier uses various contextual features such as co-occurrence words, the output of the Japanese morphological analyzer, JUMAN, the output of the Japanese parser: KNP, letter n-grams and so on. In Murata's paper, CRL uses both the Naive Bayes classifier and the Support Vector Machine classifier. Our reimplementation of CRL uses the Naive Bayes classifier with the following similar contextual feature sets (a few exceptions can be used only in SENSE-VAL2 tasks).

> **Letter N-grams**
> > Unigram, Bigram and Trigram surrounding the target word.

> **JUMAN**
> > The analysis provided by JUMAN on words surrounding the target word including POSs, word class, word sub class, and several features.

KNP

> The analysis provided by KNP parser including the type of clause containing the target word and information of modification (dependency) relation with the target word.

Co-occurrence Words

> The analysis provided by JUMAN on co-occurrence words including POSs, word class, word sub class, and several features.

noACD (no associated word)

> This method uses not only the CRL's various features but also the relevant words extracted from the Mainichi newspaper corpus. However, this method uses only the label words (stimulus words in ACD) as search queries and so does not use any associated word. By using label words to extract relevant words, this system can use the most relevant words.

ACD (with associated words)

> This is the proposed system. This system also uses CRL's various features and relevant words extracted from the Mainichi newspaper corpus. To extract relevant words, this system use both the label words (stimulus word) and associated words as the search queries as described in Section 3. Thus, this system uses fewer relevant words than the noACD system, but the extracted relevant words that are used are expected to be more closely related to each class label (context) and less noisy than those of the noACD system.

ACD-for-WSD (with associated words)

> This system uses almost the same method as ACD (with associated words) but uses ACD-for-WSD in place of ACD. By using ACD-for-WSD, specialized for the WSD task, the accuracy of the system is expected to be improved.

4.2. Naive Bayes method

We used the Naive Bayes method (NB) to evaluate CRL and our proposals. Let $T = \{t_1, t_2, \ldots, t_n\}$ represent class labels of target words and $F = \{f_1, f_2, \ldots, f_m\}$ be a set of features. In the classification process, the Naive Bayes classifier tries to determine the correct label that maximizes $P(t_i|F)$, the probability of class t_i given feature set F, $1 \leq i \leq n$. Assuming the independence of features, the classification procedure can be formulated as:

$$\hat{t_i} = \arg \max_{t_i} \left(\ln(P(t_i)) + \sum_{j=1}^{n} \ln(P(f_j|t_i)) \right) \quad (1)$$

$$P(f_j|t_i)) \approx \frac{freq(f_j|t_i) + e * freq(t_i)}{freq(t_i) + e * freq(t_i)} \quad (2)$$

where $P(t_i)$, $P(f_i|t_i)$ and $P(f_i)$ are estimated using maximum likelihood methods and $freq(x)$ is a function that counts the frequency of x. To avoid the influence of zero frequency problems, we used the above formula and set e to 0.0001.

Method	CRL's features	Search query words
Baseline (most freq sense)	-	-
CRL(Murata, 2003)	USE	-
noACD	USE	label words
ACD	USE	label and associated words in ACD
ACD-for-WSD	USE	label and associated words in ACD-for-WSD

Table 1: Summary of Methods.

Method	Mainichi	Aozora
Baseline (most freq sense)	0.36	0.25
CRL(Murata, 2003)	0.67	0.37
noACD (no associated word)	0.57	0.27
ACD (with associated words)	*0.69*	*0.69*
ACD-for-WSD (with associated words)	*0.70*	*0.69*

Table 2: Accuracy rates of each WSD method for the two different corpora (Mainichi newspaper and Aozora Bunko). Mainichi accuracy was achieved by using 10-fold cross validation. Aozora accuracy was achieved by the NB classifier trained on the Mainichi newspaper corpus (497 sentences).

5. Evaluation Results and Conclusion

Our results are summarized in Table 2. As can be seen, the proposed methods, ACD (with associated words) and ACD-for-WSD (with associated words), perform better than CRL. A comparison against noACD (no associated word) shows the effectiveness of using associated words to extract relevant words. Though the noACD method extracted more relevant words than ACD and ACD-for-WSD, both proposed methods perform better.

Furthermore, CRL demonstrated remarkably low accuracy against the Aozora Bunko corpus test set. On the other hand, the proposed methods offered higher accuracy. This is evidence that our methods are superior in terms of effectiveness and versatility to the conventional methods examined.

6. Acknowledgements

7. References

Eneko Agirre and Philip Edmonds. 2006. *Word Sense Disambiguation: Algorithms and Applications (Text, Speech and Language Technology)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Satanjeev Banerjee and Ted Pedersen. 2002. An adapted lesk algorithm for word sense disambiguation using wordnet. In *Proceedings of CICLing 2002*, pages 136–145, London, UK. Springer-Verlag.

Sadao Kurohashi and Kiyoaki Shirai. 2001. Senseval-2 japanese tasks(in japansese). volume 101, pages 1–8. The Institute of Electronics, Information and Communication Engineers.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of SIGDOC 1986*, pages 24–26. ACM New York, NY, USA.

Masaki Murata, Masao Uchiyama, Kiyotaka Uchimoto, Qing Ma, and Hitoshi Isahara. 2003. Crl at japanese dictionary based task of senseval-2(in japanese). volume 10, pages 115–133.

Jun Okamoto and Shun Ishizaki. 2001. Construction of associative concept dictionary with distance information, and comparison with electronic concept dictionary (in japanese). *Journal of Natural Language Processing*, 8(4):37.

Jun Okamoto and Shun Ishizaki. 2007. Word sense disambiguation on contextual dynamic network using associative concept dictionary. In *Proceedings of PACLING 2007*, pages 93–100.

Kiyoaki Shirai and Tsunekazu Yagi. 2004. Learning a robust word sense disambiguation model using hypernyms in definition sentences. In *Proceedings of Coling 2004*, pages 917–923, Geneva, Switzerland, Aug 23–Aug 27. COLING.

Dongqiang Yang and David M. W. Powers. 2006. Word sense disambiguation using lexical cohesion in the context. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 929–936, Sydney, Australia, July. Association for Computational Linguistics.

David Yarowsky. 1992. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of COLING 1992*, pages 454–460. Association for Computational Linguistics Morristown, NJ, USA.