In Proceedings of NLPRS'97 pp. 201-206
1997, Dec.

An Empirical Comparison of Recent Corpus-Based Word Sense Disambiguation Techniques

Atsushi Fujii, Takenobu Tokunaga, Hozumi Tanaka
Department of Computer Science, Tokyo Institute of Technology
2-12-1 Oookayama Meguroku Tokyo 152, JAPAN
{fujii,take,tanaka}@cs.titech.ac.jp

Abstract

This paper describes a comparative evaluation of recent corpus-based word sense disambiguation techniques, focusing around ten Japanese verbs. To be able to generalize the results of our experiments, we used both a correct/incorrect binary judgement and a scaled acceptability factor as evaluation criteria. Through our evaluation, we found that the similarity-based method relying on a hand-crafted thesauri generally outperformed other methods.

1 Introduction

This paper describes an extensive comparative evaluation of recent word sense disambiguation (WSD) techniques, which represent a potentially crucial component of numerous NLP applications. We currently focus on the sense disambiguation of Japanese verbs, for example, the following input sentence containing the polysemous verb *tsukau*:

kodomo ga	*kozukai wo*	*tsukau.*
(children-NOM)	(allowance-ACC)	(?)

In Japanese, each verb complement consists of a noun phrase (case filler) and a case-marking suffix (case marker), for example *ga* (nominative), *ni* (dative) or *wo* (accusative). The "EDR" Japanese machine readable dictionary (Japan Electronic Dictionary Research Institute, 1995) defines multiple senses for the verb *tsukau*, a sample of which are "to employ", "to operate" and "to spend". From among these candidates, one may notice that the correct interpretation of *tsukau* in the above input is "to spend".

Reflecting the growing utilization of machine readable texts, a number of corpus-based WSD techniques have recently been proposed. These techniques use a training example set, in which polysemous words are annotated with their correct senses (correct word senses are usually determined under supervision by human experts). Figure 1 classifies the different WSD approaches focused on in this paper. The first approach, in what we shall call the word-based method, simply relies on collocational statistics, disregarding syntactic relations. This approach is used in many WSD techniques for noun sense disambiguation (some of which are reviewed, for example, by Charniak (1993)). The second approach is sensitive to the syntactic and semantic content of complements of a polysemous verb, based on the intuitively feasible assumption that the sense of a verb is likely to be dependent on its syntactically governing complements. We can further subdivide the syntactic approach into three different subapproaches. The first approach is the rule-based method, which uses a thesaurus to (automatically) identify appropriate semantic classes as selectional restrictions for each verb complement (Resnik, 1993; Ribas, 1995). The second approach can be called the Naive-Bayes method (Ng, 1997; Pedersen et al., 1997), which estimates the probability that a polysemous verb input takes each sense, and selects the verb sense with the highest probability. The third approach, in what we shall call the similarity-based method, relies on the similarity between an input and supervised examples. This approach is found in a number of verb sense disambiguation techniques (Fujii et al., 1996a; Kurohashi and Nagao, 1994; Li et al., 1995; Uramoto, 1994).

Section 2 elaborates on the candidate methods given in figure 1, and section 3 compares them by way of experiments. Discussion is added in section 4, followed by the conclusion.

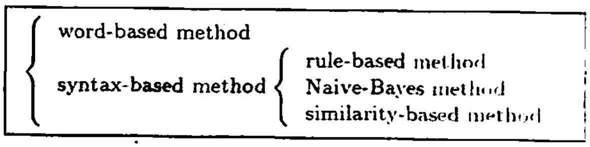

Figure 1: The different corpus-based word sense disambiguation methods

2 The different methodologies

2.1 Word-based method

The word-based method selects the verb sense based on the association degree between collocating words contained in the input, and each verb sense. In this method, the input is morphologically analyzed (lexically segmented and tagged with parts-of-speech), and collocational data determined only for nouns, because functional words such as case markers are generally more noisy than informative. Thereafter, the verb sense with the maximal association degree is selected (equation (1)).

$$\arg\max_{s} \sum_{w \in input} P(s|w) \qquad (1)$$

Here, $P(s|w)$ denotes the conditional probability that verb sense s occurs, given word w[1]. This factor is calculated based on training data, prior to the disambiguation process.

2.2 Syntax-based method

In the syntax-based method, syntactic analysis is performed subsequent to morphological analysis to extract the verb-complement structure from the input. This analysis is especially poignant when the input comprises a complex sentence. Based on the extracted syntactic structure, we first discard verb sense candidates with case frames not corresponding to the obligatory case content of the input. The case frame of each verb sense is given as the case pattern of supervised examples associated with that verb sense. Those discarded candidates are not considered in the ensuing disambiguation process. In the following sections, we explain three submethods.

2.2.1 Rule-based method

In the rule-based method, the selectional restrictions are represented by thesaurus classes, and allow only those nouns dominated by the given class in the thesaurus structure as verb complements. In order to identify appropriate thesaurus classes, we used the association measure proposed by Resnik (1993), which computes the information-theoretic association degree between case fillers and thesaurus classes, for each verb sense (equation (2)).

$$A(s,c,r) = P(r|s,c) \cdot \log \frac{P(r|s,c)}{P(r|c)} \qquad (2)$$

Here, $A(s,c,r)$ is the association degree between verb sense s and class r (selectional restriction candidate) with respect to case c. $P(r|s,c)$ is the conditional probability that a case filler example associated with case c of sense s is dominated by class r in the thesaurus. $P(r|c)$ is the conditional probability that a case filler example for case c (disregarding verb sense) is dominated by class r. Each probability is estimated based on training data. We used the semantic classes defined in the *Bunruigoihyo* thesaurus (National Language Research Institute, 1996). In practice, every r whose association degree is above a certain threshold is chosen as a selectional restriction (Ribas. 1995)[2].

2.2.2 Naive-Bayes method

The Naive-Bayes method (Ng, 1997: Pedersen et al., 1997) assumes that each case filler included in the input is conditionally independent of other case fillers: the system approximates the probability that an input x takes a verb sense s ($P(s|x)$), simply by computing the product of the probability that each verb sense s takes x_c as a case filler for case c. The verb sense with maximal probability is then selected as the interpretation (equation (3)).

$$\begin{aligned} \arg\max_{s} P(s|x) &= \arg\max_{s} \frac{P(s) \cdot P(x|s)}{P(x)} \\ &= \arg\max_{s} P(s) \cdot P(x|s) \\ &\simeq \arg\max_{s} P(s) \prod_{c} P(x_c|s) \end{aligned} \qquad (3)$$

Here, $P(x_c|s)$ is the probability that a case filler associated with sense s for case c in the training data is x_c. We estimated $P(s)$ based on the distribution of the verb senses in the training data. In practice, data sparseness leads to not all case fillers x_c appearing in the database, and as such, we generalize each x_c into a 5 digit semantic class defined based on the *Bunruigoihyo* thesaurus (National Language Research Institute. 1996).

2.2.3 Similarity-based method

The similarity-based method, which takes after the nearest neighbor method, searches the training example set for the most semantically similar example to the input. Thereafter, the polysemous verb included in the input is disambiguated by superimposing the sense of the verb appearing in the example. Figure 2 depicts a general schema for computation of the similarity between the input and each example. In this figure, x denotes the input, and e denotes an example associated with

[1] Note that other factors can also be used for this computation, such as mutual information between each collocating word and verb sense. However, our preliminary experiment showed that this factor did not improve on the system performance.

[2] We conducted several trials prior to the actual experiment, to determine the optimal threshold value.

verb sense s. x_c and e_c denote the case fillers marked with case c, in x and e, respectively. The similarity between x and e ($sim(x, e)$) is computed by summing the similarity between the input case filler and the example case filler for each case, as given in equation (4).

$$sim(x, e) = \sum_{c \in x} sim(x_c, e_c) \qquad (4)$$

One may notice that the critical content of this method is the method of computing $sim(x_c, e_c)$, that is, the similarity between the case fillers x_c and e_c. We explain two methods for this computation below.

Vector space model The first approach is based on statistical modeling. We adopted one typical implementation called the "vector space model" (VSM), which has a long history of application in information retrieval (IR) and text categorization (TC) tasks (Frakes and Baeza-Yates, 1992). In the case of IR/TC, VSM is used to compute the similarity between documents, which is represented by a vector comprising statistical term weights of content words in a document. Similarly, in our case, each noun is represented by a vector comprising term weights, although term weights are calculated for verbs co-occurring with the noun. We used TF·IDF for term weighting, and the similarity between two nouns is computed as the cosine of the angle between their associated vectors.

Co-occurrence data was extracted from the RWC text base RWC-DB-TEXT-95-1 (Real World Computing Partnership, 1995). The RWC text base consists of 4 years worth of Mainichi Shimbun newspaper articles (Mainichi Shimbun, 1991-1994), which have been automatically annotated with morphological tags. The total number of co-occurrences was 419,132.

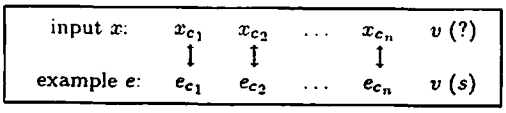

Figure 2: Similarity computation between the input and examples, under the similarity-based method

Hand-crafted thesaurus A number of WSD techniques (Fujii et al., 1996a; Kurohashi and Nagao, 1994; Li et al., 1995; Uramoto, 1994) use existing hand-crafted thesauri (for example, Roget's thesaurus (Chapman, 1984), WordNet (Miller et al., 1993) or *Bunruigoihyo* (National Language Research Institute, 1996)) for the similarity computation, based on the intuitively feasible assumption that words located near each other within the structure of a thesaurus have similar meaning. Therefore, the similarity between two given words is represented by the length of the path between them in the thesaurus structure. We experimentally used two different thesauri. One is the *Bunruigoihyo* thesaurus (National Language Research Institute, 1996), which has been commonly used for much Japanese NLP research. The other is the NTT thesaurus (Ikehara et al., 1997), which associates each entry word with multiple concepts while the *Bunruigoihyo* thesaurus fundamentally associates each word with a single concept. When nouns are associated with multiple concepts, we determine the similarity between each combination of senses associated with the given nouns, and take the maximal similarity (equation (5)).

$$sim(n_1, n_2) = \max_{c_1, c_2} sim(c_1, c_2) \qquad (5)$$

Here, c_1 and c_2 denote senses associated with nouns n_1 and n_2, respectively.

3 Comparative experiments

3.1 Methodology

We collected sentences (as test/training data) from the EDR Japanese corpus (Japan Electronic Dictionary Research Institute, 1995), originally produced from news articles. The EDR corpus provides sense information for each word, based on the EDR dictionary, and we used this as a means of checking the interpretation. Our derived corpus contains ten verbs frequently appearing in the EDR corpus, with the total number of sentences being 10,880. For each of the ten verbs, we conducted 4-fold cross validation: that is, we divided the corpus into four equal parts, and conducted four trials, in each of which a different one of the four parts was used as test data and the remaining parts were used as training data. We compared the following six methods (hereafter, we shall use the notation "VSM", "NTT" and "BGH", for the vector space model, the use of the NTT thesaurus and the use of the *Bunruigoihyo* thesaurus, respectively):

(1) lower bound method, in which the system systematically chooses the verb sense appearing most frequently in the database (Gale et al., 1992),

(2) word-based method,

(3) rule-based method,

(4) Naive-Bayes method,

(5) similarity-based method (VSM),

(6) similarity-based method (NTT),

(7) similarity-based method (BGH).

For morph/syntax analysis, we used the Japanese "QJP" parser (Kameda, 1996), which was used to conduct syntactic analysis for methods (3) to (7), and morphological analysis for methods (2) to (7). In Japanese, complements of a verb are not always provided, and often omitted if they are easily predictable (based on human judgment) from the context. In such a situation, methods (3) to (7) simply use method (1). In method (3), when all verb senses are rejected by selectional restrictions, method (1) is used additionally.

3.2 Evaluation criterion

Instead of the conventional binary (correct/incorrect) judgment, a number of researchers have advocated an evaluation criterion which is based on the semantic similarity between word sense candidates (Fujii et al., 1997; Lin, 1997; Resnik and Yarowsky, 1997). To exemplify this notion, let us take the Japanese polysemous verb ***tsukau*** again, which has multiple senses in EDR, such as "to employ", "to operate", "to spend" and "to use MATERIAL". These senses are associated with the EDR thesaurus (Japan Electronic Dictionary Research Institute, 1995). Figure 3 shows a fragment of the thesaurus, in which an oval denotes a word sense. As with most thesauri, the length of the path between two word senses can be seen as the relative semantic distance between them. As one can see, the verb sense "to spend" is physically closer to "to use MATERIAL" than to "to employ" or "to operate", in structure. In fact, these two proximal verb senses can be merged into one common category, that is, "to use up". Furthermore, they can be merged with "to operate" to form "to use PHYSICAL OBJECT", as distinct from the remaining sense of "to employ (HUMAN/CONCEPT)". Let the correct sense of an input ***tsukau*** be "to use MATERIAL", and assume the system incorrectly outputs "to spend". In this case, the error would be more acceptable than outputing "to employ", which is totally different to the correct interpretation of "to use MATERIAL". Therefore, we should allot the system a scaled acceptability factor instead of a score of zero. Note that the binary judgment simply scores the system zero, irrespective of the extent to which the error is acceptable given a particular practical application.

We formalized the scaled acceptability factor ("acceptability", hereafter) as in equation (6).

$$A(x,s) = \left(\frac{MAXLEN - EDR(x,s)}{MAXLEN}\right)^{\alpha} \qquad (6)$$

Here, x and s are the system's interpretation and the correct answer, respectively, and $A(x,s)$ is the acceptability of the given x and s. $EDR(x,s)$ represents the length of the path between x and s in the EDR thesaurus. $MAXLEN$ is the maximum length of the path between senses associated with an input verb. For example, $MAXLEN = 7$ in figure 3 (the length of the path between "to operate" and "to employ"). α is a parametric constant, which acts to control the reduction factor of $A(x,s)$ for incorrect interpretations. With a larger α, $A(x,s)$ becomes smaller for incorrect interpretations, and becomes closer to the binary judgment. One can notice that the acceptability ranges from 0 (where x and s are the most dissimilar verb senses) to 1 (x and s are identical).

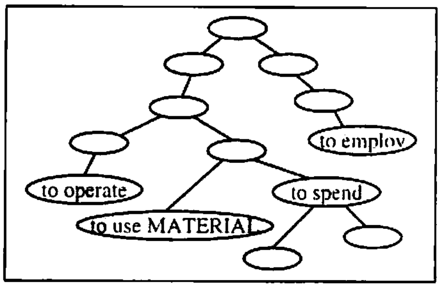

Figure 3: A fragment of the EDR thesaurus including senses related to the Japanese verb *tsukau*

Table 1 shows the acceptability of each method for three different α values and the accuracy based on the binary judgement (the "binary" column). In this case, the accuracy is the ratio of the number of correct interpretations, to the number of outputs. One can see that method (7) generally outperformed the other methods, although methods (6) and (7) are comparable in performance. Note that given smaller values for α (that is, by relaxing the evaluation criterion), method (3) outperformed the lower bound (method (1)). It should also be noted that since the similarity-based method dynamically searches the example data set, the execution time can be prohibitive when compared with the word-based and Naive-Bayes methods. However, an existing sampling method (Fujii et al., 1996b), which selects an informative smaller-sized example set, is expected to overcome this problem.

Table 1: The acceptability of each method, with varying values for α

	binary	$\alpha = 1$	$\alpha = 2$	$\alpha = 0.5$
(1)	.449	.540	.634	.738
(2)	.501	.671	.772	.861
(3)	.409	.601	.723	.831
(4)	.569	.711	.798	.877
(5)	.591	.740	.822	.893
(6)	.615	.754	.832	.898
(7)	.626	.761	.837	.901

4 Discussion

Is full parsing needed? As shown in table 1, syntax-based methods (both the Naive-Bayes and similarity-based methods) improved on the performance for the word-based methods. However, given the fact that automatic grammar acquisition does not seem to be advanced enough to be practical, syntax-based methods require the manual construction of a grammar for syntactic analysis. In other words, we should investigate the trade-off between the improvement of verb sense disambiguation and the overhead required for syntactic analysis.

One may notice that partial parsing with simple heuristics can be enough for verb sense disambiguation because we only need complements of the target verb, rather than a full syntactic analysis. For partial parsing, we used the two simple heuristics given below:

- each complement (noun + case marker) is associated to the predicate of highest proximity,
- complements containing the genitive case marker *no* are not considered because they can constitute either possessive or nominative case markers, and are thus confusing.

We found that the accuracy of method (7) (the use of the *Bunruigoihyo* thesaurus) combined with partial parsing was 62.0%, which is almost equal to the accuracy with full parsing in table 1 (acceptability showed the same tendency). Therefore, we can reduce the overhead required for syntactic analysis, without degrading the system performance.

Is human knowledge needed? One may argue that given sufficient statistics, the vector space model should outperform hand-crafted thesauri, in other words, human lexicographers' knowledge is no longer needed. We investigate this prediction in table 2, which shows the the relation between the frequency of nouns appearing in the co-occurrence data extracted from the RWC text base (see section 2.2.3) and the accuracy of verb sense disambiguation, in which the "frequency" entry denotes the threshold of the frequency of nouns. The "coverage" entry denotes the ratio between the number of inputs including at least one noun with frequency over a given threshold, and the total number of inputs. The last two entries show the accuracy with different similarity measures, for each coverage. Surprisingly, not only the accuracy of VSM but also the accuracy of NTT and BGH increased as the threshold of the frequency increased, and VSM did not outperform BGH for any of the thresholds. The acceptability (with $\alpha = 1$) for each coverage is also shown in parentheses, which shows the same tendency as for the accuracy. We could assume frequently appearing nouns are used so commonly that even human lexicographers can reasonably define the similarity between them in a thesaurus. In addition, nouns which frequently appear in the co-occurrence data also appear in the training data, and therefore they provide the maximal similarity (that is, "exact matching") independent of which similarity measure is used. We would like to note that human knowledge is useful in the task of word sense disambiguation, as with other NLP research (Klavans and Resnik, 1996).

Table 2: The frequency of nouns and resultant accuracy of verb sense disambiguation

frequency	≥100	≥1000	≥10000
coverage	73.9%	58.2%	16.8%
VSM	.650 (.843)	.699 (.863)	.734 (.877)
NTT	.674 (.853)	.707 (.867)	.736 (.877)
BGH	.687 (.859)	.723 (.874)	.747 (.882)

Incorporation of contextual constraint A number of researchers have pointed out that words tend to maintain the same sense within a given context (Nasukawa, 1993; Yarowsky, 1995). In other words, when the same verb appears multiply in the same context, we assume that it takes the same sense. The crucial issue is then which verb sense to take if each verb occurrence is interpreted with different senses by a dedicated method (for example, those reviewed in this paper). We newly introduce a method to propagate contextual constraint through the degree of interpretation certainty. Let us assume that a given input includes two distinct verbs of common lexical content, and that one of them is interpreted with greater certainty than the other. In such a case, we superimpose the interpretation with higher certainty onto the interpretation with lower certainty. Computation of the degree of certainty is performed using the method proposed by Fujii et al (1996b) (see their paper for details). We collected sentences in which a polysemous verb appears more than once, from the corpus used in our experiments above[3]. The number of derived sentences was 462 (out of 10,880), which means the applicability of this method is relatively small. However, this method improved the accuracy from 60.4% to 64.1%, when used in conjunction with method (7).

5 Conclusion

In this paper, we reviewed recent approaches for verb sense disambiguation and compared them

[3] We limited the range of context to one sentence because the EDR corpus does not provide wider contextual information, such as paragraph boundaries and sentence genres.

by way of experiments. To be able to generalize the result of our experiments, we used both a correct/incorrect binary judgement and a scaled acceptability factor as evaluation criterion. Through our evaluation, we verified that the following items: (a) syntactic relations between a target verb and its complements improved on the performance for simple word-based method, without the considerable overhead for syntactic analysis, (b) the similarity-based method combined with handcrafted thesauri outperformed the lower bound method to a larger degree than other methods and (c) our prototype method to propagate contextual constraints further improved on the performance of the similarity-based method.

Acknowledgments

The authors would like to thank Masayuki Kameda (RICOH Co., Ltd., Japan) for his support with the QJP parser, Timothy Baldwin (TITECH, Japan) for his comments on the earlier version of this paper, and NTT staff (NTT, Japan) for their support with the NTT thesaurus.

References

Robert L. Chapman. 1984. *Roget's International Thesaurus (Fourth Edition).* Harper and Row.

Eugene Charniak. 1993. *Statistical Language Learning.* MIT Press.

William B. Frakes and Ricardo Baeza-Yates. 1992. *Information Retrieval: Data Structure & Algorithms.* PTR Prentice-Hall.

Atsushi Fujii, Kentaro Inui, Takenobu Tokunaga, and Hozumi Tanaka. 1996a. To what extent does case contribute to verb sense disambiguation? In *Proceedings of COLING*, pages 59–64.

Atsushi Fujii, Kentaro Inui, Takenobu Tokunaga, and Hozumi Tanaka. 1996b. Selective sampling of effective example sentence sets for word sense disambiguation. In *Proceedings of the Fourth Workshop on Very Large Corpora*, pages 56–69.

Atsushi Fujii, Kentaro Inui, Takenobu Tokunaga, and Hozumi Tanaka. 1997. Evaluation of word sense disambiguation systems. In *Proceedings of the third Annual Meetings of the Association for Natural Language Processing*, pages 305–308. (In Japanese).

William Gale, Kenneth Ward Church, and David Yarowsky. 1992. Estimating upper and lower bounds on the performance of word-sense disambiguation programs. In *Proceedings of ACL*, pages 249–256.

Satoshi Ikehara et al. 1997. *A Japanese Lexicon.* Iwanami Shoten. (In Japanese).

Japan Electronic Dictionary Research Institute. 1995. EDR electronic dictionary technical guide. (In Japanese).

Masayuki Kameda. 1996. A portable & quick Japanese parser : QJP. In *Proceedings of COLING*, pages 616–621.

Judith Klavans and Philip Resnik, editors. 1996. *The Balancing Act: Combining Symbolic and Statistical Approaches to Language.* MIT press.

Sadao Kurohashi and Makoto Nagao. 1994. A method of case structure analysis for Japanese sentences based on examples in case frame dictionary. *IEICE Transactions on Information and Systems.* E77-D(2):227–239.

Xiaobin Li, Stan Szpakowicz, and Stan Matwin. 1995. A WordNet-based algorithm for word sense disambiguation. In *Proceedings of IJCAI*, pages 1368–1374.

Dekang Lin. 1997. Using syntactic dependency as local context to resolve word sense ambiguity. In *Proceedings of ACL*, pages 64–71.

Mainichi Shimbun. 1991-1994. Mainichi shimbun CD-ROM '91-'94. (In Japanese).

George A. Miller et al. 1993. Five papers on WordNet. Technical report, Cognitive Science Laboratory, Princeton University.

Tetsuya Nasukawa. 1993. Discourse constraint in computer manuals. In *Proceedings of TMI*, pages 183–194.

National Language Research Institute. 1996. *Bunruigoihyo.* Shuei publisher, revised and enlarged edition. (In Japanese).

Hwee Tou Ng. 1997. Exemplar-based word sense disambiguation: Some recent improvements. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing.*

Ted Pedersen, Rebecca Bruce, and Janyce Wiebe. 1997. Sequential model selection for word sense disambiguation. In *Proceedings of ANLP.*

Real World Computing Partnership. 1995. RWC text database. (In Japanese).

Philip Resnik and David Yarowsky. 1997. A perspective on word sense disambiguation methods and their evaluation. In *Proceedings of ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?*

Philip Stuart Resnik. 1993. *Selection and Information: A Class-Based Approach to Lexical Relationships.* Ph.D. thesis, University of Pennsylvania.

Francesc Ribas. 1995. On learning more appropriate selectional restrictions. In *Proceedings of EACL.*

Naohiko Uramoto. 1994. Example-based word-sense disambiguation. *IEICE Transactions on Information and Systems*, E77-D(2):240–246.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of ACL*, pages 189–196.