人工知能学会 全国大会 1997 pp. 350−353

事例に基づく動詞多義性解消における事例の類似度計算について

Similarity Computation in Verb Sense Disambiguation

藤井 敦　　徳永健伸　　田中穂積
Atsushi Fujii　　Takenobu Tokunaga　　Hozumi Tanaka
東京工業大学大学院情報理工学研究科
Department of Computer Science, Tokyo Institute of Technology

Abstract: This paper describes a comparative evaluation of recent corpus-based word sense disambiguation techniques, focusing around ten Japanese verbs. The basis of this task is the computation of the similarity between a given input and examples which have been annotated with verb sense, and we tentatively compare different methods using four different types of information: morphological content, syntactic structure, semantic similarity, and contextual constraint. We also introduce a new method for propagating contextual constraints. Through our experiments and discussion, we found the performance of word sense disambiguation improved more as we used more information, and the improvement in accuracy was a maximum of about 30%.

1 Introduction

This paper describes a comparative evaluation of recent word sense disambiguation (WSD) techniques, which represent a crucial component of numerous NLP applications. We currently focus on the sense disambiguation of Japanese verbs, for example, the following input sentence containing the sense ambiguous 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 [3] defines multiple senses for the verb *tsukau*, a sample of which are "to employ", "to operate" and "to spend". Among these candidates, one may notice that the correct interpretation of *tsukau* in the above input is "to spend". Note that the task of disambiguation as discussed in this paper can be termed *categorization* because the plausible verb sense is selected from predefined candidates.

Reflecting the growing utilization of machine readable texts, a number of corpus-based WSD techniques have recently been proposed. The crucial content of these methods is the computation of the scored similarity between a given input and an example sentence set already annotated with verb sense (mostly annotated by human experts). We shall call such a set of examples a "database". The verb sense with maximal similarity score is then selected as the interpretation of the input verb. In the following, we tentatively classify existing WSD techniques from the viewpoint of processing complexity (in descending order):

1. word-based method: the input and examples in the database are simply morphologically analyzed, and the similarity between them is computed based on the words contained in them,
2. syntax-based method: besides morphological content, the syntactic structures of the input and examples are also used in the similarity computation,
3. thesaurus-based method: the similarity between the input and examples is computed by use of semantic resources, that is, hand-crafted thesauri,
4. context-based method: verbs appearing in the same context (i.e. a sentence or paragraph) commonly share the same sense.

It can be assumed that the performance of word sense disambiguation will improve as the processing mechanism becomes more complicated. On the other hand, one may argue that higher-level processing poses a considerable overhead, and therefore, there is a trade-off between the performance and such processing complexity (for example, syntactic analysis involves the overhead for constructing a base grammer). To minimize this overhead, in the case of syntactic analysis, we introduce simple heuristics to take the place of a rich grammar. The effectivity of this method as compared to an existing parser is discussed through experiments.

In section 2, we elaborate on the above four different methodologies for similarity computation. We then compare their effectivity by way of experiments in section 3. Discussion is added in section 4, followed by our conclusion.

2 The different methodologies

2.1 Word-based method

In the word-based method, the database contains a set of words collocating with each verb sense. In real terms, we only used collocational information for nouns, because functional words such as case markers are generally more noisy than informative. Based on the database, the statistical factor for each collocating

藤井 敦 〒 251 目黒区大岡山 2-1-12 Tel (03)5734-2831 Fax (03)5734-2915 e-mail: fujii@cs.titech.ac.jp

word and verb sense, that is, the degree of association between them, is calculated prior to sense disambiguation. In regard to the score for each verb sense, a commonly proposed implementation is used, in which the score is computed by summing the statistical factor of each collocating word which appears in a given input [18], as in equation (1).

$$Score(s) = \sum_{w \in input} A(s,w) \tag{1}$$

Here, $A(s,w)$ denotes the degree of association between sense s and each collocating word w.

In the following two paragraphs, we explain two types of statistical factors which were used in our experiments.

Mutual information The notion of *mutual information* is used in much NLP research for estimating the degree of association between two given terms. In our implementation, $A(s,w)$ is expressed by equation (2), where $f(s,w)$ denotes the frequency of w collocating with sense s, and $f(s)$ and $f(w)$ denote the frequency of w and s, respectively. All these factors are calculated based solely on the database.

$$A(s,w) = \frac{f(s,w)}{f(s) \cdot f(w)} \tag{2}$$

Conditional probability Another implementation calculates $A(s,w)$ as the probability that s occurs, when w occurs [15]. This is expressed by equation (3), closely resembling that for mutual information.

$$A(s,w) = \frac{f(s,w)}{f(w)} \tag{3}$$

2.2 Syntax-based method

Syntactic analysis, in which the verb-complement structure is extracted from a given input, is especially poignant when the input comprises a complex sentence. To achieve this process, we have two candidate methods. On the one hand, full parsing with rich grammar rules is ideal. On the other hand, partial parsing with simple heuristics can be preferable because (a) the manual construction of a grammar is expensive, (b) automatic grammar acquisition does not seem to be advanced enough to be practical, and (c) we only need complements of the target verb, rather than a full syntactic analysis.

To conduct full parsing, we experimentally used the Japanese "QJP" parser [8]. We also used this parser as the morphological analyzer for all four WSD approaches focused on in this paper. In regard to partial parsing, we introduced 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.

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 a verb sense is given as the case pattern of each example associated with a given verb sense in the database. Those discarded candidates are not considered in the following process. We then compute the similarity between an input and each example based on their case structures. Figure 1 depicts a general schema for this notion, in which x denotes an input, and e denotes an example associated with verb sense s in the database. x_c and e_c denote the case fillers marked with case c, in x and e, respectively. Intuitively speaking, two case structures are more similar if they share more case fillers. However, since each x_c does not always appear in the database due to data sparseness, we need to employ a smoothing technique. Note that, in the word-based method, data sparseness can be avoided to a large extent because we can use not only complements of a target verb, but also every collocating word in a given input.

$$\boxed{\begin{array}{cccccc} \text{input } x\text{:} & x_{c_1} & x_{c_2} & \cdots & x_{c_n} & v\ (?) \\ & \updownarrow & \updownarrow & & \updownarrow & \\ \text{example } e\text{:} & e_{c_1} & e_{c_2} & \cdots & e_{c_n} & v\ (s) \end{array}}$$

Figure 1: Similarity computation between the input and examples under the syntax-based approach

There are a number of statistical models for word similarity measurement, but the model in widest usage is the "vector space model" (VSM) [2, 7, 16]. We used this word similarity measure for the smoothing of input case fillers. In this model, each noun n is represented by a vector comprising statistical co-occurrence factors. This can be expressed by equation (4), where $\vec{n}$ is the vector for the noun in question, and terms t_i are the co-occurrence statistics of n and each co-occurring verb.

$$\vec{n} = < t_1,\ t_2,\ \ldots,\ t_i,\ \ldots > \tag{4}$$

Co-occurrence data was extracted from the RWC text base RWC-DB-TEXT-95-1 [13]. This text base consists of 4 years worth of Mainichi Shimbun [14] newspaper articles, which have been automatically annotated with morphological tags. The total morpheme content is about 100 million. Instead of conducting full parsing on the text, several heuristics were used in order to obtain dependencies between complements (noun + case marker) and verbs, in the form of tuple-based templates $<n,c,v>$. In regard to t_i, we used the notion of TF·IDF [4], in which t_i is calculated as in equation (5), where $f(<n,c,v>)$ is the frequency of the tuple $<n,c,v>$, $f(<c,v>)$ is the frequency of tuple $<c,v>$, and N is the total number of tuples within the overall co-occurrence data.

$$t_i = f(<n,c,v>) \cdot \log \frac{N}{f(<c,v>)} \tag{5}$$

We then compute the similarity between nouns x_c and e_c by the cosine of the angle between the two vectors $\vec{x_c}$ and $\vec{e_c}$. This is realized by equation (6).

$$sim(x_c, e_c) = \frac{\vec{x_c} \cdot \vec{e_c}}{|\vec{x_c}||\vec{e_c}|} \tag{6}$$

The *total* similarity between an input and examples is computed by summing the similarity between the input case filler and the example case filler for each case, as in equation (7).

$$sim(x,e) = \sum_{c \in x} sim(x_c, e_c) \tag{7}$$

Here, $sim(x,e)$ denotes the similarity between the input x and example e, and $sim(x_c,e_c)$ denotes the similarity between the case fillers x_c and e_c, which can be measured based on table 1.

Finally, based on previously proposed methods [9, 17], we compute the score for sense s by way of equation (8)

$$Score(s) = \sum_{c \in x} \max_{e_c \in \mathcal{E}_s} sim(x_c, e_c) \qquad (8)$$

2.3 Thesaurus-based method

With regard to the smoothing of case fillers, we can use semantic resources, that is, hand-crafted thesauri (for example, Roget's thesaurus [1], WordNet [10] or *Bunruigoihyo* [12]), based on the intuitively feasible assumption that words located near each other within the structure of a thesaurus have similar meaning. For our experiments, we used the Japanese *Bunruigoihyo* thesaurus, which is commonly used in much NLP research, and applied the similarity metric proposed by Kurohashi et al., as shown in table 1. We then compute the score for a verb sense by replacing equation (7) with the similarity given by table 1.

Table 1: The relation between the length of the path between two nouns n_1 and n_2 in the *Bunruigoihyo* thesaurus ($len(n_1, n_2)$) and their relative similarity ($sim(n_1, n_2)$)

$len(n_1, n_2)$	0	2	4	6	8	10	12	14
$sim(n_1, n_2)$	12	11	10	9	8	7	5	0

2.4 Context-based method

A number of researchers have pointed out that words tend to maintain the same sense within a given context [11, 19]. 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 lower-level similarity computation (such as the thesaurus-based method). We newly introduce a method to avoid this problem, in which we compute the degree of certainty of verb sense disambiguation for each occurrence, and select the verb sense with maximal certainty degree. With regard to computation of the certainty degree, we simply used our previously proposed technique [5], as shown in equation (9), where $C(x)$ is the interpretation certainty of an example x. $Score_1(x)$ and $Score_2(x)$ are the highest and second highest scores for x, respectively. λ, which ranges from 0 to 1, is a parametric constant (we set $\lambda = 0.5$).

$$C(x) = \lambda \cdot Score_1(x) + (1 - \lambda) \cdot (Score_1(x) - Score_2(x)) \qquad (9)$$

3 Comparative experiments

We collected sentences (as test/training data) from the EDR Japanese corpus [3] (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 interpretations. Our derived corpus consists of 10880 sentences containing ten verbs frequently appearing in the EDR corpus. 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 (the database). In this experiment, we compared the following methods (these methods are in ascending order in terms of processing complexity, excepting methods (2) and (3), which are of equivalent complexity):

(1) lower bound: a naive method, in which the system systematically chooses the verb sense appearing most frequently in the database [6].
(2) word-based method (mutual information),
(3) word-based method (conditional probability),
(4) syntax-based method (partial parsing),
(5) syntax-based method (full parsing),
(6) use of the *Bunruigoihyo* thesaurus (partial parsing),
(7) use of the *Bunruigoihyo* thesaurus (full parsing),

In Japanese, complements of a verb are not always provided because they are often omitted if they are easily predictable (based on human judgment) from the context. In such a situation, methods (4) to (7) simply use method (1). Table 2 shows the accuracy of each method, given that the accuracy is the ratio of the number of correct interpretations, to the number of outputs. It should be noted that according to our preliminary observation, the EDR corpus contains a number of sense tagging errors, and we assume this is why the accuracy of each approach was generally not acceptable. However, all methods except (2) outperformed the lower bound results produced by method (1), and as we assumed, the accuracy became greater as the processing mechanism became more complicated. The improvement in accuracy between methods (2) and (7) was more than 30%. We assume that mutual information gives lower values for frequently appearing verb senses, and therefore those verb senses are rarely selected as the interpretation. Surprisingly, comparing methods (4) and (5) (or methods (6) and (7)), the two parsing techniques did not yield a significant difference in terms of the accuracy. Therefore, we can improve on the performance of verb sense disambiguation without the considerable overhead for syntactic analysis.

Table 2: The accuracy of each method (%)

(1)	(2)	(3)	(4)	(5)	(6)	(7)
44.9	29.2	50.1	59.0	59.1	62.0	62.6

With regard to the context-based method, we limited the range of context to one sentence because the EDR corpus does not provide wider contextual information, such as paragraph boundaries and genres. We collected sentences containing multiple instances of a single verb, the number of which was 462. For these sentences, our context-based method improved the accuracy from 60.4% to 64.1%, when used in conjunction with method (7).

4 Discussion

One may argue that given sufficient statistics, the smoothing method provided by the vector space model should outperform use of the *Bunruigoihyo* thesaurus. We investigated this prediction in table 3, which shows the the relation between the frequency of nouns appearing in the co-occurrence data 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 the *Bunruigoihyo* thesaurus increased as the threshold of the frequency increased, and VSM did not outperform the *Bunruigoihyo* thesaurus for any of the thresholds. We could assume that (a) nouns which frequently appear in the co-occurrence data also appear in the database, and therefore they provide the maximum similarity (that is, "exact matching") independent of which similarity measure is used, and (b) frequently appearing nouns are used so commonly that even human lexicographers can reasonably define the similarity between them in a thesaurus. Putting this result aside, we do not believe our co-occurrence data was insufficient because it was taken from *four* years worth of newspaper articles.

Table 3: The frequency of nouns and resultant accuracy of verb sense disambiguation

frequency	≥100	≥1000	≥10000
coverage (%)	73.9	58.2	16.8
Bunruigoihyo	68.7	72.3	74.7
VSM	65.0	69.9	73.4

In the syntax-based method, we only considered noun phrases containing case marker for complements of a verb. However, there are other types of postpositions such as *wa* (topic marker) and *mo* ("also"). Kurohashi et al. proposed a way of modelling these verb complements by matching them to complements followed by *ga*, *ni* or *wo* based on the similarity between respective case fillers [9]. If this process is carried out successfully, the similarity between an input and examples is expected to be more reliable. We applied this technique to method (7) in section 3, and its accuracy was 54.0%. As far as the corpus we used was concerned, we conclude that further analysis for case matching is needed.

5 Conclusion

In this paper, we classified recent approaches for verb sense disambiguation based on processing complexity, and compared them by way of experiments. As far as we are aware, this represents the largest-scale attempt to compare the various methods. We also introduced a new method of propagating contextual constraints base on the certainty degree. Through evaluation, we concluded that as the disambiguation process becomes more complicated (ranging from the simple use of morphological analysis to intergrating the use of contextual constraint), we gained higher accuracy of word sense disambiguation. In addition, we introduced simple heuristics for syntactic analysis, which contributed to accuracy gain as much as an existing syntactic analyzer. Future work will include the evaluation of the effectivity of further processing, including discourse analysis [11, 19] and ellipsis/anaphora analysis to recover complements, which we could not conduct using the EDR corpus.

Acknowledgments

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

References

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

[2] Eugene Charniak. *Statistical Language Learning*. MIT Press, 1993.

[3] EDR. *EDR Electronic Dictionary Technical Guide*, 1995. (In Japanese).

[4] William B. Frankes and Ricardo Baeza-Yates. *Information Retrieval: Data Structure & Algorithms*. PTR Prentice-Hall, 1992.

[5] Atsushi Fujii, Kentaro Inui, Takenobu Tokunaga, and Hozumi Tanaka. Selective sampling of effective example sentence sets for word sense disambiguation. In *Proceedings of the Fourth Workshop on Very Large Corpora*, pp. 56–69, 1996. http://xxx.lanl.gov/ps/cmp-lg/9702010.

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

[7] Donald Hindle. Noun classification from predicate-argument structures. In *Proceedings of ACL*, pp. 268–275, 1990.

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

[9] Sadao Kurohashi and Makoto Nagao. A method of case structure analysis for Japanese sentences based on examples in case frame dictionary. *IEICE TRANSACTIONS on Information and Systems*, Vol. E77-D, No. 2, pp. 227–239, 1994.

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

[11] Tetsuya Nasukawa. Discourse constraint in computer manuals. In *Proceedings of TMI*, pp. 183–194, 1993.

[12] National Language Research Institute. *Bunruigoihyo*, revised and enlarged edition, 1996. (In Japanese).

[13] Real World Computing Partnership. RWC text database. http://www.rwcp.or.jp/wswg.html, 1995.

[14] Mainichi Shimbun. Mainichi shimbun CD-ROM '91-'94, 1991-1994.

[15] Takenobu Tokunaga and Makoto Iwayama. Text categorization based on weighted inverse document frequency. *Information Processing Society of Japan, SIGNL*, Vol. 94, No. 100, pp. 33–40, 1994.

[16] Takenobu Tokunaga, Makoto Iwayama, and Hozumi Tanaka. Automatic thesaurus construction based on grammatical relations. In *Proceedings of IJCAI*, pp. 1308–1313, 1995.

[17] Naohiko Uramoto. Example-based word-sense disambiguation. *IEICE TRANSACTIONS on Information and Systems*, Vol. E77-D, No. 2, pp. 240–246, 1994.

[18] David Yarowsky. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of COLING*, pp. 454–460, 1992.

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