Natural Language Processing Pacific Rim Symposium '95

Word Sense Disambiguation by Marker Passing on Very Large Semantic Networks

OKUMURA Manabu,
School of Information Science,
Japan Advanced Institute of
Science and Technology
Tatsunokuchi, Ishikawa 923-12, Japan
E-mail: oku@jaist.ac.jp

MATSUNAGA Yuukichi
Ohtori Signal&Telecommunication Section,
JR-WEST
Ohtori-minami machi 3-333,
Sakai-city, Osaka, Japan
E-mail: CXP05343@niftyserve.or.jp

Abstract

Some works that use machine readable dictionaries for word sense disambiguation compute the number of shared words in definition. The first problem of these works is that their methods fail to take into account less immediate relationships between word senses. Secondly, their methods do not take into account the degree of importance of words in definition. In this paper, to cope with the above two problems, we present a method for word sense disambiguation with very large semantic networks(VLSN) that computes the score which depends on the number of shared words in definition. We experimentally show to what extent the words in common should be searched for. We also show the scoring measure that takes into account the degree of importance of words in definition.

1 Introduction

Word sense disambiguation is a difficult problem in natural language analysis(Hirst, 1987). Since senses of words in a sentence tend to be semantically related to each other, word sense ambiguity can be resolved by selecting the sense of a word which is semantically related to senses of other words in a sentence. Some works that use machine readable dictionaries for word sense disambiguation can be considered to be based on this idea. They compute the number of shared words in definition(Lesk, 1986; Cowie et al., 1992; Demetriou, 1993). In these methods, two word senses are thought to be semantically more related if they have more words in common in their definitions. Therefore, the sense of a word with the greatest number of shared words with senses of other words is selected as the correct one.

The first problem of these works is that their methods fail to take into account less immediate relationships between word senses. Even if two word senses have no words in common in their definitions, they might have one in the definitions of words that appear in their definitions. Therefore, it is important to decide to what extent the words in common should be searched for in definitions.

Secondly, their methods do not take into account the degree of importance of words in definition. Even if word senses have a word in common in their definitions, this overlap might have little significance if the word appears quite frequently in definitions. Therefore, it is necessary to give the word in common a weight that depends on its degree of importance. Moreover, because of the discrete measure for word sense selection, their methods cannot determine the correct sense if multiple senses have the same number of overlaps.

In this paper, to cope with the above problems, we present a method for word sense disambiguation with very large semantic networks(VLSN) that computes the score which depends on the number of shared words in definition. We automatically build VLSN with definitions in a machine readable version of Japanese dictionary. It consists of about 400,000 nodes and about 1.5 million links. Because of this very largeness, it is implemented on massively parallel computers. We experimentally show to what extent the words in common should be searched for. We also show the scoring measure that takes into account the degree of importance of words in definition.

(Veronis and Ide, 1990) also indicated the above points and proposed a method with very large neural networks. Their method, in fact, succeeds to take into account less immediate relationships between word senses, but their framework does not give an answer to the above question about the appropriate extent of search. Moreover, their method is not evaluated fully with regard to whether weighting the overlaps is effective in word sense disambiguation.

In section two, we describe our method for word sense disambiguation with VLSN and show the scoring measure that takes into account the degree of importance of words in definition. In section three, we describe the experiments with our VLSN and evaluate our word sense disambiguation method.

2 Word Sense Disambiguation with VLSN

First, we describe how VLSN can be built from machine readable dictionaries(MRD). We assume that there are semantic relations between a word and the words which are used to define it. The connections in the network indicate the relations. A word node is connected by links to sense nodes that represent the different possible senses for the word. Each sense node is connected to word nodes that represent the words in the definition of the sense. By repeating this connection, MRD produces the network in which word nodes and sense nodes are connected alternately. Our VLSN is implemented on the connection machine CM-5 which has 64 processors.

2.1 Marker Passing as Overlap Detection

In this section, we explain that marker passing in the network can find an overlap in definition. Consider, for example, the following imaginary definitions of words W1 and Wi.

```
W1:
  S11 : ...
  S12 : ...
  ...
  S1m1: ..., Wx, ...
...
Wi:
  Si1 : ..., Wx, ...
  ...
  Simi: ...
```

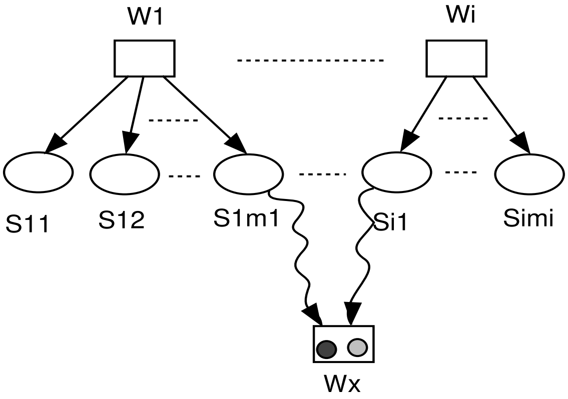

Figure 1: The detection of the words in common

If the words W1 and Wi appear in a sentence, the pair of word senses S1m1 and Si1 is selected as their senses, since only they have the word Wx in common in their definitions.

Assuming semantic networks, we can find a path between multiple nodes if markers from the two nodes are collided at a node(Charniak, 1986; Norvig, 1989; Alshawi, 1987). As shown in figure 1 [1], the node(Wx) where the markers from two nodes are collided can be considered as the overlap in their definitions. Therefore, detecting an overlap between multiple sense nodes can be realized by finding paths between the nodes by marker passing. Marker passing based overlap detection is widely applicable in that it can find the overlap not only in their direct definitions, but in the definitions of words that appear in their definitions by searching further along links in the network.

2.2 Computing the Score

In the previous approaches to word sense disambiguation by overlap detection, each word that is found as an overlap is equally counted as one. However, some words appear quite frequently in definitions and they are thought to have less significance as the overlap. Moreover, because of the discrete measure for word sense selection, the previous methods cannot determine the correct sense if multiple senses have the same number of overlaps.

---

[1] Boxes and ovals represent word and sense nodes respectively in the figures of this paper.

Besides, now we search for an overlap not only in the direct definition but in the whole network, we should regard the overlap that is found in the direct definition as more important than the one that is found by searching further in the network. Moreover, the sense ambiguity of words in definition might degrade the reliability of the overlap that is found by searching further in the network.

Taking into account the above points, we compute the score for word sense selection which depends on the number of overlaps.

The score is computed as the sum of the activation values of all the overlap nodes. If the number of overlaps increases, the score gets greater. The activation value of an overlap node is computed from the activation value of the markers which collide at the node. Since the overlap which is found at the earlier step of marker passing should be more important and get more activation value than the one at the later step, the activation value of the marker should be decayed whenever it goes through a node. The degree of the decay of the activation value depends on the importance of the node where the marker goes through. Therefore, the activation value is multiplied by the degree of importance of the node which ranges between 0 and 1 whenever the marker goes through the node [2]. Finally, the path which includes the node with the greatest score is selected as the most appropriate one and the word senses which are included in the path are selected as the correct ones.

2.2.1 The Importance of the Nodes

Words in definition have the degree of importance. The words which appear quite frequently in definitions tend to be not a keyword. Therefore, we compute the degree of importance of word nodes by the frequency of the word nodes in definitions. The fragment of the network in figure 2 shows that the word W appears in n definitions of different word senses. In this case, the importance of the word W is tentatively defined in the following equation:

$$I_W(n) = \frac{1}{n+A} + B,$$

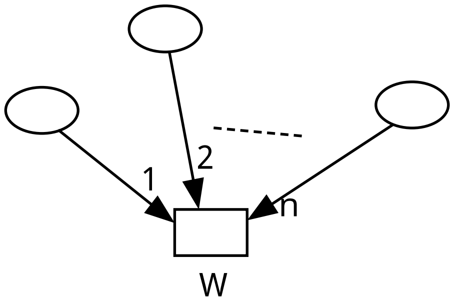

Figure 2: The Importance of the Word Node

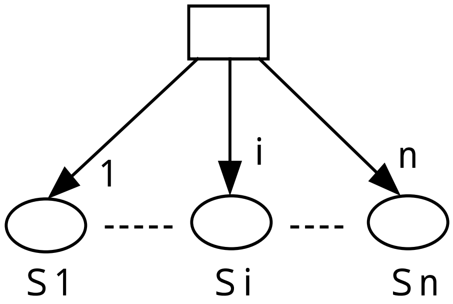

Figure 3: The Importance of the Sense Node

where A and B are parameters to be fixed experimentally. $I_W(n)$ ranges between 0 and 1.

Secondly, since words in definition have multiple senses and the path that is found might include wrong senses of the words, the reliability of the overlap that is found might degrade. Since this degree of the degrade can be considered to depend on the number of senses of words in definition, we define the degree of importance of sense nodes as the number of their competitive word senses. As shown in figure 3, if the word has n senses, the importance of the node is tentatively defined as follows:

$$I_{S_i}(n) = \frac{1}{n}.$$

Since the frequency of each word sense is not given in our MRD, we give uniform importance to each word sense.

In the future, we will have to present the best equations for defining importance of the nodes.

2.2.2 Computing the Activation Value of the Overlap Node

The activation value of the overlap word node W(A(W)) is computed as the sum of two activation values: the activation value of the node itself(M(W)) and the activation value of the overlap nodes which were found before(N(W)). Since markers that collided at a node might collide again with markers from the other nodes, the activation value of the node represents the sum of the activation values of all the overlap nodes that were found before.

---

[2] We try to reduce the combination of senses of words incrementally during marker passing to avoid the combinatorial explosion of word sense combinations and so the scoring measure that is really used is a revision of the measure that is described here. But for the sake of simplicity, we do not explain about the revision.

Assuming $a_1, \ldots, a_n$ as the activation value of markers which collide at the node W, the activation value of the node W itself is computed as the sum of all a_i [3].

3 The Experiment

In this section, we first describe the experiment with our VLSN to fix two parameters on the importance of the node. Then, we evaluate our word sense disambiguation method with fixed parameters.

3.1 Fixing the Parameters

As we described in the last section, we introduce the notion of the importance of the node. As for the importance of word nodes, we have to fix two parameters that are included in the equation in section 2.2.1. Furthermore, as we explained before, we have to decide to what extent the words in common should be searched for. Therefore, we now have the following three parameters to be fixed:

- The maximum value of the importance of the word node(MAX_I)
- The ratio of the maximum to the minimum value of the importance of the word node($RATIO$)
- To what extent the marker should be propagated from the initial nodes($STOP$)

That $STOP$ is 1 means that the word in common is searched only in the definitions of the word senses that are tried to be disambiguated. That $STOP$ is 2 means that markers are propagated to the word nodes in the definitions of the words in the definitions of the target word senses. MAX_I and $RATIO$ are used for computing parameters A and B in the equation in section 2.2.1.

We use the following training data for fixing the parameters [4]:

(a)	(b)	(c)
50	265(5.3)	193(3.86)

(a): No. of sentences
(b): No. of words
(c): No. of ambiguous words

The number in parentheses in the table is the average number of words in a sentence. The parameters range as follows:

parameters	value range
MAX_I	0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9
$RATIO$	2,7,12,17,22,27,32,37,42
$STOP$	1,2,3,4

Since these parameters are independent of each other, the total number of combinations is 324.

The combination of parameters with the best precision will be adopted. The result of the disambiguation procedure is checked against a data manually analyzed by one of the authors, and the precision for a sentence is computed as the quotient of the number of correctly disambiguated words by the number of ambiguous words that the system can disambiguate. If the system can disambiguate no words in a sentence, the precision for the sentence is zero. If the system's output is a tie for a word, the number of correctly disambiguated words is computed as the quotient of 1 by the number of word senses that the system outputs. The average precision for the whole data of N sentences is computed from the precision for a sentence, as follows:

$$Precision = \frac{1}{N'} \sum_{i=1}^{N} Precision_i,$$

where N' is the number of sentences for which the system can output any answers.

We get the following results from the experiment:

- When $STOP$ is 1, the number of words that can be disambiguated is 63 and the number of correctly disambiguated words is 29, and the average precision is 45.2%, regardless of the other two parameters.
- When $STOP$ is 4, the increase of the number of words that can be disambiguated exceeds the increase of the number of correctly disambiguated words, therefore, the average precision becomes quite lower than when $STOP$ is 2 or 3.

---

[3] The previous approaches by overlap detection without the notion of the importance of the node can be realized by assigning the same value to all a_i.

[4] Common function words were removed from each sentence.

- When *STOP* is 2 or 3, the other parameters are fixed as follows, because of the best precision:

	STOP	*MAX_I*	*RATIO*
A	2	0.2	42
B	3	0.3	32

In the next section, we evaluate our system mainly with these combinations of parameters A and B.

3.2 The Evaluation

We evaluate our word sense disambiguation method with the combination of the parameters that was fixed in the last section. We use the following test data for the evaluation:

(a)	(b)	(c)
50	211(4.68)	177(3.93)

(a): No. of sentences
(b): No. of words
(c): No. of ambiguous words

The system's performance is shown in the next table:

	(a)	(b)	average precision(%)
A	75	32	48.2
B	88	40	52.7

(a): No. of words that can be disambiguated
(b): No. of correctly disambiguated words

In case where *MAX_I* and *RATIO* are both 1, our system can simulate the previous approaches that compute the number of overlaps without the notion of the importance of the node. The result in this case is as follows:

STOP	(a)	(b)	average precision(%)
2	150	50.5	35.5
3	161	56.5	34.9

(a): No. of words that can be disambiguated
(b): No. of correctly disambiguated words

So far, all the content words in a sentence are tried to be disambiguated all together. However, since it is rare for all the words in a sentence to be related to each other, next we evaluate the system's performance with only the words that are related to each other in a sentence. We use the same set of sentences as the above evaluation, but only the set of content words that one of the authors judges to be related to each other are extracted from them. Therefore, the data that we use for the new evaluation is as follows:

(a)	(b)	(c)
34	75(2.21)	60(1.76)

(a): No. of sentences
(b): No. of words
(c): No. of ambiguous words

The system's performance for the new data is shown in the next table:

	(a)	(b)	average precision(%)
A	28	21	79.4
B	35	23	67.5

(a): No. of words that can be disambiguated
(b): No. of correctly disambiguated words

The result in case where our system simulates the previous approaches for the new data is as follows:

STOP	(a)	(b)	average precision(%)
2	49	19.1	37.6
3	51	22.3	39.4

(a): No. of words that can be disambiguated
(b): No. of correctly disambiguated words

From these results, we can state the following points:

- In case without the notion of the importance of the node, in fact it is not an exact comparison, but the case where *STOP* is 1, that is, searching only in the direct definition, gets the best precision, compared with the case where *STOP* is 2 or 3. Therefore, the previous approaches that compute the overlaps only in the direct definition can be considered to be the best method when the notion of the importance of the node is not introduced.

- Comparing the precision of the cases where the notion of the importance of the node is introduced and is not introduced, we can claim that it is clearly better to introduce it [5]. Therefore, we think that our approach with the importance of the node is superior to the previous ones.

- As for our approach with the importance of the node, unfortunately we cannot conclude which is better between the combinations of parameters A and B, although both A and B are clearly better than when $STOP$ is 1, concerning the number of the words that can be disambiguated and the average precision.
- Analyzing the results of the two experiments where all the content words and only the related words in a sentence are respectively tried to be disambiguated, we think that our system's performance might be improved if our system is extended to select multiple paths between related word senses. Our system now takes into account only the case where the words A, B, C, and D in a sentence are all related to each other, and cannot cope with the case where only the word pairs A-C and B-D are respectively related to each other in a sentence. The latter case can be considered to be treated by selecting multiple paths.

4 Conclusion

We presented a method for word sense disambiguation with very large semantic networks(VLSN) that computes the score which depends on the number of shared words in definition. In fact our data that was used for evaluation is small, we experimentally showed to what extent the words in common should be searched for. We think it is meaningful to suggest the extent of the search when MRD is used for word sense disambiguation. Furthermore, we showed that it is better to introduce the importance of the node into word sense disambiguation with overlap detection.

Like the previous approaches(Cowie et al., 1992; Demetriou, 1993), we will have to take into account the combinatorial problem of word senses seriously, although we now cope with the problem by the naive method that was mentioned in section 2.2.

References

H. Alshawi. 1987. *Memory and Context for Language Interpretation.* Studies in Natural Language Processing. Cambridge University Press.

E. Charniak. 1986. A neat theory of marker passing. In *Proc. of the 5th National Conference on Artificial Intelligence*, pages 584–588.

J. Cowie, J. Guthrie, and L. Guthrie. 1992. Lexical disambiguation using simulated annealing. In *Proc. of the 14th International Conference on Computational Linguistics*, pages 359–365.

G.C. Demetriou. 1993. Lexical disambiguation using constraint handling in prolog(chip). In *Proc. of the 6th European Chapter of the Association for Computational Linguistics*, pages 431–436.

G. Hirst. 1987. *Semantic Interpretation and the Resolution of Ambiguity.* Studies in Natural Language Processing. Cambridge University Press.

M. Lesk. 1986. Automated sense disambiguation using machine-readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proc. of the ACM SIGDOC Conference*, pages 24–26.

P. Norvig. 1989. Marker passing as a weak method for text inferencing. *Cognitive Science*, 13(4):569–620.

J. Veronis and N.M. Ide. 1990. Word sense disambiguation with very large neural networks extracted from machine readable dictionaries. In *Proc. of the 13th International Conference on Computational Linguistics*, pages 389–394.

---

[5] It is true that the method without the importance handles more sentences, but we think this broader coverage causes the poorer precision.