Word Sense Disambiguation and Text Segmentation Based on Lexical Cohesion

OKUMURA Manabu, HONDA Takeo
School of Information Science,
Japan Advanced Institute of Science and Technology
(Tatsunokuchi, Ishikawa 923-12 Japan)
e-mail:{oku,honda}@jaist.ac.jp

Abstract

In this paper, we describe how word sense ambiguity can be resolved with the aid of lexical cohesion. By checking lexical cohesion between the current word and lexical chains in the order of the salience, in tandem with generation of lexical chains, we realize incremental word sense disambiguation based on contextual information that lexical chains reveal. Next, we describe how segment boundaries of a text can be determined with the aid of lexical cohesion. We can measure the plausibility of each point in the text as a segment boundary by computing a degree of agreement of the start and end points of lexical chains.

1 Introduction

A text is not a mere set of unrelated sentences. Rather, sentences in a text are about the same thing and connected to each other[10]. *Cohesion* and *coherence* are said to contribute to such connection of the sentences. While coherence is a semantic relationship and needs computationally expensive processing for identification, cohesion is a surface relationship among words in a text and more accessible than coherence. Cohesion is roughly classified into *reference*[1], *conjunction*, and *lexical cohesion*[2].

Except conjunction that explicitly indicates the relationship between sentences, the other two classes are considered to be similar in that the relationship between sentences is indicated by two semantically same(or related) words. But lexical cohesion is far easier to identify than reference because both words in lexical cohesion relation appear in a text while one word in reference relation is a pronoun or elided and has less information to infer the other word in the relation automatically.

Based on this observation, we use lexical cohesion as a linguistic device for discourse analysis. We call a sequence of words which are in lexical cohesion relation with each other a *lexical chain* like [10]. Lexical chains tend to indicate portions of a text that form a semantic unit. And so various lexical chains tend to appear in a text corresponding to the change of the topic. Therefore,

1. lexical chains provide a local context to aid in the resolution of word sense ambiguity;
2. lexical chains provide a clue for the determination of segment boundaries of the text[10].

In this paper, we first describe how word sense ambiguity can be resolved with the aid of lexical cohesion. During the process of generating lexical chains incrementally, they are recorded in a register in the order of the salience. The *salience* of lexical chains is based on their recency and length. Since the more salient lexical chain represents the nearby local context, by checking lexical cohesion between the current word and lexical chains in the order of the salience, in tandem with generation of lexical chains, we realize incremental word sense disambiguation based on contextual information that lexical chains reveal.

Next, we describe how segment boundaries of a text can be determined with the aid of lexical cohesion. Since the start and end points of lexical chains in the text tend to indicate the start and end points of the segment, we can measure the plausibility of each point in the text as a segment boundary by computing a degree of agreement of the start and end points of lexical chains.

[1] Reference by pronouns and ellipsis in Halliday and Hasan's classification[3] are included here.

[2] Reference by full NPs, substitution and lexical cohesion in Halliday and Hasan's classification are included here.

Morris and Hirst[10] pointed out the above two importance of lexical cohesion for discourse analysis and presented a way of computing lexical chains by using Roget's International Thesaurus[15]. However, in spite of their mention to the importance, they did not present the way of word sense disambiguation based on lexical cohesion and they only showed the correspondences between lexical chains and segment boundaries by their intuitive analysis.

McRoy's work[8] can be considered as the one that uses the information of lexical cohesion for word sense disambiguation, but her method does not take into account the necessity to arrange lexical chains dynamically. Moreover, her word sense disambiguation method based on lexical cohesion is not evaluated fully.

In section two we outline what lexical cohesion is. In section three we explain the way of incremental generation of lexical chains in tandem with word sense disambiguation and describe the result of the evaluation of our disambiguation method. In section four we explain the measure of the plausibility of segment boundaries and describe the result of the evaluation of our measure.

2 Lexical Cohesion

Consider the following example, which is the English translation of the fragment of one of Japanese texts that we use for the experiment later.

> In the universe that continues expanding, a number of stars have appeared and disappeared again and again. And about ten billion years after the birth of the universe, in the same way as the other stars, a primitive galaxy was formed with the primitive sun as the center.

Words {universe, star, universe, star, galaxy, sun} seem to be semantically same or related to each other and they are included in the same category in Roget's International Thesaurus. Like Morris and Hirst, we compute such sequences of related words(lexical chains) by using a thesaurus as the knowledge base to take into account not only the repetition of the same word but the use of superordinates, subordinates, and synonyms.

We use a Japanese thesaurus 'Bunrui-goihyo'[1]. Bunrui-goihyo has a similar organization to Roget's: it consists of 798 categories and has a hierarchical structure above this level. For each word, a list of category numbers which corresponds to its multiple word senses is given. We count a sequence of words which are included in the same category as a lexical chain. It might be clear that this task is computationally trivial. Note that we regard only a sequence of words in the same category as a lexical chain, rather than using the complete Morris and Hirst's framework with five types of thesaural relations.

The word sense of a word can be determined in its context. For example, in the context {universe, star, universe, star, galaxy, sun}, the word 'earth' has a 'planet' sense, not a 'ground' one. As clear from this example, lexical chains can be used as a contextual aid to resolve word sense ambiguity[10]. In the generation process of lexical chains, by choosing the lexical chain that the current word is added to, its word sense is determined. Thus, we regard word sense disambiguation as selecting the most likely category number of the thesaurus, as similar to [16].

Earlier we proposed incremental disambiguation method that uses intrasentential information, such as selectional restrictions and case frames[12]. In the next section, we describe incremental disambiguation method that uses lexical chains as intersentential(contextual) information.

3 Generation of Lexical Chains

In the last section, we showed that lexical chains can play a role of local context. However, multiple lexical chains might cooccur in portions of a text and they might vary in their plausibility as local context. For this reason, for lexical chains to function truly as local context, it is necessary to arrange them in the order of the salience that indicates the degree of the plausibility. We base the salience on the following two factors: the recency and the length. The more recently updated chains are considered to be the more activated context in the neighborhood and are given more salience. The longer chains are considered to be more about the topic in the neighborhood and are given more salience.

By checking lexical cohesion between the current word and lexical chains in the order of the salience, the lexical chain that is selected to add the current word determines its word sense and plays a role of local context.

Based on this idea, incremental generation of lexical chains realizes incremental word sense disambiguation using contextual information that lexical chains reveal. During the generation of lexical chains, their salience is also incrementally updated. We think incremental disambiguation[9] is a better strategy, because a combinatorial explosion of the number of total ambiguities might occur if ambiguity is not resolved as early as possible during the analytical process. Moreover, incremental word sense disambiguation is indispensable during the generation of lexical chains if lexical chains are used for incremental analysis, because the word sense ambiguity might cause many undesirable lexical chains and they might degrade the performance of the analysis(in this case, the disambiguation itself).

3.1 The Algorithm

First of all, a Japanese text is automatically segmented into a sequence of words by the morphological analysis[11]. From the result of the morphological analysis, candidate words are selected to include in lexical chains. We consider only nouns, verbs, and adjectives, with some exceptions such as nouns in adverbial use and verbs in postpositional use.

Next lexical chains are formed. Lexical cohesion among candidate words inside a sentence is first checked by using the thesaurus. Here the word sense of the current word might be determined. This preference for lexical cohesion inside a sentence over the intersentential one reflects our observation that the former might be tighter.

After the analysis inside a sentence, candidate words are tried to be added to one of the lexical chains that are recorded in the register in the order of the above salience. The first chain that the current word has the lexical cohesion relation is selected. The salience of the selected lexical chain gets higher and then the arrangement in the register is updated.

Here not only the word sense ambiguity of the current word is resolved but the word sense of the ambiguous words in the selected lexical chain can also be determined. Because the lexical chain gets higher salience, other word senses of the ambiguous words in the lexical chain which correspond to other lexical chains can be rejected. Therefore, lexical chains can be used not only as prior context but also later context for word sense disambiguation.

If a candidate word can not be added to the existing lexical chain, new lexical chains for each word sense are recorded in the register.

As clear from the algorithm, rather than the truly incremental method where the register of lexical chains is updated word by word in a sentence, we adopt the incremental method where updates are performed at the end of each sentence because we regard intrasentential information as more important.

The process of word sense disambiguation using lexical chains is illustrated in Figure 1. The most salient lexical chain is located at the top in the register. In the initial state the word *W1* remains ambiguous. When the current unambiguous word *W2* is added, the chain *b* is selected(top-left). The chain *b* becomes the most salient(top-right). Here the word sense ambiguity of the word *W1* in the chain *b* is resolved(bottom-left). If the word to be added is ambiguous(*W3*), the word sense corresponding to the more salient lexical chain(*ID31*) is selected(bottom-right).

3.2 The Evaluation

We apply the algorithm to five texts. Table 1 shows the system's performance.

The 'correctness' of the disambiguation is judged by one of the authors. The system's performance is computed as the quotient of the number of correctly disambiguated words by the number of ambiguous words minus the number of wrongly segmented words(morphological analysis errors)[3].

Words that remain ambiguous are those that do not form any lexical chains with other words. Except by the errors in the morphological analysis, most of the errors in the disambiguation are caused by being dragged into the wrong context.

The average performance is 63.4 %. We think the system's performance is promising for the following reasons:

1. Lexical cohesion is not the only knowledge source for word sense disambiguation and proves to be useful at least as a source supplementary to our earlier framework that used case frames[12].

2. In fact, higher performance is reported in [16], that uses broader context acquired by

[3] The accuracy of the morphological analysis will be improved by adding new word entries or the like.

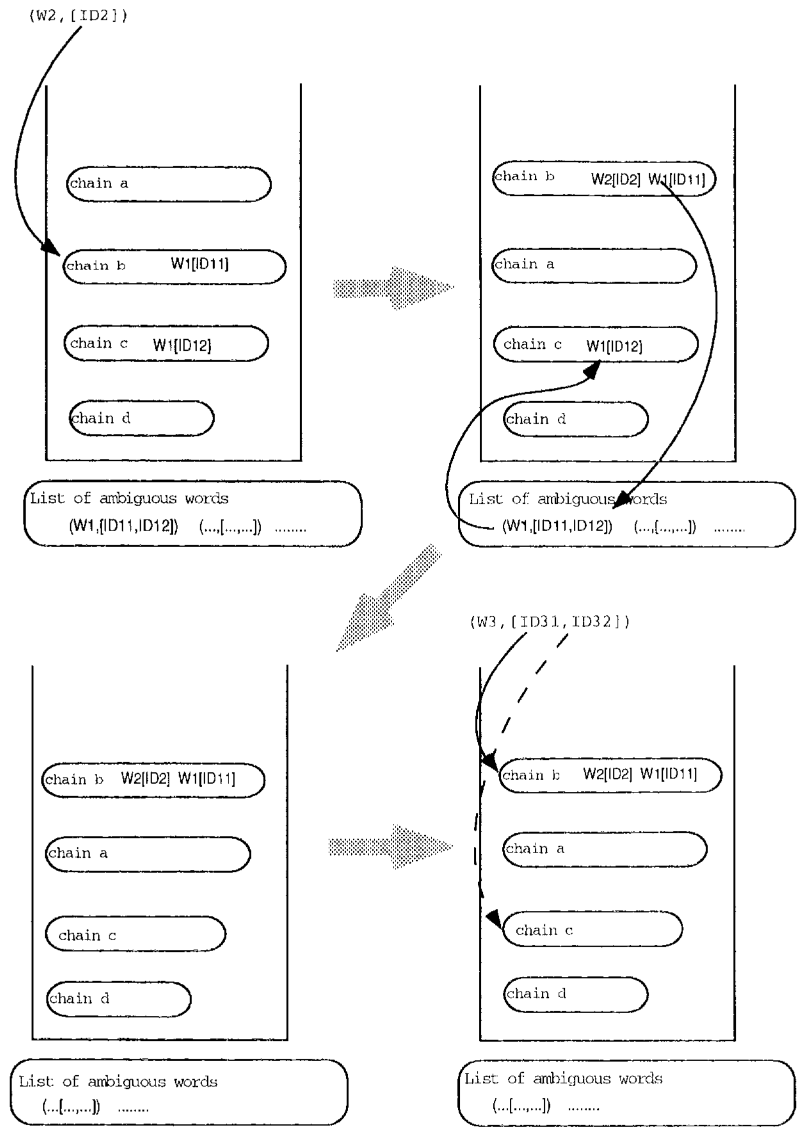

Figure 1: The process of word sense disambiguation

text	number of sentences	number of candidate words	number of ambiguous words	number of words that remain ambiguous	number of correctly disambiguated words	system's performance (%)
No.1	41	481	166	7	126	87.5
No.2	26	197	71	13	32	51.6
No.3	24	212	57	12	34	64.2
No.4	38	433	123	19	71	60.1
No.5	24	163	82	11	42	53.8

Table 1: The performance for the disambiguation

training on large corpora, but our method can attain such tolerable level of performance without any training.

However, our salience of lexical chains is, of course, rather naive and must be refined by using other kinds of information, such as Japanese topical marker 'wa'.

4 Text Segmentation by Lexical Chains

The second importance of lexical chains is that they provide a clue for the determination of segment boundaries. Certain spans of sentences in a text form semantic units and are usually called segments. It is crucial to identify the segment boundaries as a first step to construct the structure of a text[2].

4.1 The Measure for Segment Boundaries

When a portion of a text forms a semantic unit, there is a tendency for related words to be used. Therefore, if lexical chains can be found, they will tend to indicate the segment boundaries of the text. When a lexical chain ends, there is a tendency for a segment to end. If a new chain begins, this might be an indication that a new segment has begun[10]. Taking into account this correspondence of lexical chain boundaries to segment boundaries, we measure the plausibility of each point in the text as a segment boundary: for each point between sentences n and $n+1$ (where n ranges from 1 to the number of sentences in the text minus 1), compute the sum of the number of lexical chains that end at the sentence n and the number of lexical chains that begin at the sentence $n+1$. We call this naive measure of a degree of agreement of the start and end points of lexical chains $w(n, n+1)$ *boundary strength* like [14]. The points in the text are selected in the order of boundary strength as candidates of segment boundaries.

Consider for example the five lexical chains in the imaginary text that consists of 24 sentences in Figure 2. In this text, the boundary strength can be computed as follows: $w(3,4) = 1, w(7,8) = 1, w(9,10) = 1, w(13,14) = 3, \ldots$

```
   chains   |           text
            |          1         2
 start-end  | 123456789012345678901234
 ( 1 - 24)  | ************************
 ( 4 - 13)  |    **********
 (14 - 16)  |              ***
 ( 8 -  9)  |        **
 (14 - 18)  |              *****
```

Figure 2: Lexical chains in the text

4.2 The Evaluation

We try to segment the texts in section 3.2 and apply the above measure to the lexical chains that were formed. We pick out three texts(No.3,4,5), which are from the exam questions of the Japanese language that ask us to partition the texts into a given number of segments. The system's performance is judged by the comparison with segment boundaries marked as an attached model answer. Two more texts(No.6,7) from the questions are also tried to be segmented.

Here we do not take into account the information of paragraph boundaries, such as the indentation, at all in the following reasons:

- Because our texts are from the exam questions, many of them have no marks of paragraph boundaries;
- In case of Japanese, it is pointed out that paragraph and segment boundaries do not always coincide with each other[13].

Table 2 shows the performance in case where the system generates the given number of segment boundaries[4] in the order of the strength. From Table 2, we can compute the system's marks as an examinee in the test that consists of these five questions. Table 3 shows the performance in case where segment boundaries are generated down to half of the maximum strength. The metrics that we use for the evaluation are as follows: *Recall* is the quotient of the number of correctly identified boundaries by the total number of correct boundaries. *Precision* is the quotient of the number of correctly identified boundaries by the number of generated boundaries.

We think the poor result for the text No.5 might be caused by the difficulty of the text

[4] The number of boundaries to be given is the number of segments given in the question minus 1.

text	given number of boundaries	number of correct boundaries
No.3	1	1
No.4	6	3
No.5	1	0
No.6	4	3
No.7	3	1

Table 2: The performance for the segmentation(1)

text	number of generated boundaries	number of correct boundaries	rec.	prec.
No.3	3	1	1	0.33
No.4	10	3	0.5	0.30
No.5	3	0	0	0
No.6	7	3	0.75	0.43
No.7	5	1	0.33	0.20

Table 3: The performance for the segmentation(2)

itself because it is written by one of the most difficult writers in Japan, KOBAYASHI Hideo. Table 2 shows that our system gets 8(1+3+3+1)/15(1+6+1+4+3)= 53 % in the test. From Table 3, the average recall and precision rates are 0.52 and 0.25 respectively. Of course these results are unsatisfactory, but we think this measure for segment boundaries is promising and useful as a preliminary one.

Since lexical chains are considered to be different in their degree of contribution to segment boundaries, we are now refining the measure by taking into account their importance. We base the importance of lexical chains on the following two factors:

1. The lexical chains that include more words with topical marker 'wa' get more importance.

2. The longer lexical chains tend to represent a semantic unit and get more importance.

The start and end points of the more important lexical chains can get the more boundary strength. This refinement of the measure is in the process and yields a certain extent of improvement of the system's performance.

Moreover, this evaluation method is not necessarily adequate since partitioning into a larger number of smaller segments might be possible and be necessary for the given texts. And so we will have to consider the evaluation method that the agreement with human subjects is tested in future. However, since human subjects do not always agree with each other on segmentation[6, 4, 14], our evaluation method using the texts in the questions with model answers is considered to be a good simplification.

Several other methods to text segmentation have been proposed. Kozima[7] and Youmans[17] proposed statistical measures(they are named LCP and VMP respectively), which indicate the plausibility of text points as a segment boundary. Their hills or valleys tend to indicate segment boundaries. However, they only showed the correlation between their measures and segment boundaries by their intuitive analysis of few sample texts, and so we cannot compare our system's and their performance precisely.

Hearst[5] independently proposes a similar measure for text segmentation and evaluates the performance of her method with precision and recall rates. However, her segmentation method depends heavily on the information of paragraph boundaries and always partitions a text at the points of paragraph boundaries.

5 Conclusion

We showed that lexical cohesion can be used as a knowledge source for word sense disambiguation and text segmentation. We think our method is promising, although only partially successful results can be obtained in the experiments so far. Here we reported some preliminary positive results and made some suggestions for how to improve the method in future. The improvement of the method is now under way.

In addition, because computation of lexical chains depends completely on the thesaurus used, we think the comparison among the results by different thesauri would be insightful and are now planning. It it also necessary to incorporate other textual information, such as clue words, which can be computationally accessible to improve the performance.

References

[1] *Bunrui-Goihyo.* Shuei Shuppan., 1964. in Japanese.

[2] B.J. Grosz and C.L. Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204, 1986.

[3] H. A. K. Halliday and R. Hassan. *Cohesion in English.* Longman, 1976.

[4] M.A. Hearst. Texttiling: A quantitative approach to discourse segmentation. Technical Report 93/24, University of California, Berkeley, 1993.

[5] M.A. Hearst. Multi-paragraph segmentation of expository texts. Technical Report 94/790, University of California, Berkeley, 1994.

[6] J. Hirschberg and B. Grosz. Intonational features of local and global discourse structure. In *Proc. of the Darpa Workshop on Speech and Natural Language*, pages 441–446, 1992.

[7] H. Kozima. Text segmentation based on similarity between words. In *Proc. of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 286–288, 1993.

[8] S.W. McRoy. Using multiple knowledge sources for word sense discrimination. *Computational Linguistics*, 18(1):1–30, 1992.

[9] C.S. Mellish. *Computer Interpretation of Natural Language Descriptions.* Ellis Horwood, 1985.

[10] J. Morris and G. Hirst. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1):21–48, 1991.

[11] Nagao Lab., Kyoto University. *Japanese Morphological Analysis System JUMAN Manual Version 1.0*, 1993. in Japanese.

[12] M. Okumura and H. Tanaka. Towards incremental disambiguation with a generalized discrimination network. In *Proc. of the 8th National Conference on Artificial Intelligence*, pages 990–995, 1990.

[13] T. Ookuma. Gengo tan'i toshite no bunshou. *Nihongo gaku*, 11(4):20–25, 1992. in Japanese.

[14] R.J. Passonneau. Intention-based segmentation: Human reliability and correlation with linguistic cues. In *Proc. of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 148–155, 1993.

[15] P. Roget. *Roget's International Thesaurus, Fourth Edition.* Harper and Row Publishers Inc., 1977.

[16] D. Yarowsky. Word-sense disambiguation using statistical models of roget's categories trained on large corpora. In *Proc. of the 14th International Conference on Computational Linguistics*, pages 454–460, 1992.

[17] G. Youmans. A new tool for discourse analysis: The vocabulary-management profile. *Language*, 67:763–789, 1991.