Structural Ambiguity Resolution Using Three Word Dependency Relations

Eduardo de Paiva Alves, Haodong Wu and Teiji Furugori
Department of Computer Science
University of Electro-Communications

1 Introduction

Structural ambiguity is a major obstacle in building sound systems in natural language processing. It appears in constructions such as nominal compounds, coordinated structures and prepositional phrases:

(1) *hydrogen* ion exchange
(2) *cardiac* and vascular patients
(3) I bought flowers *with Jane.*

Often the correct syntactic structure can be determined by using the lexical preferences of the words involved. In the examples above, we know that *hydrogen* prefers *ion* to *exchange*, *cardiac* prefers *vascular* to *patient*, and *with Jane* prefers *bought* to *flowers.* In studying syntactic disambiguation, many researchers have used word co-occurrences in large corpora as an indicator that shows lexical preferences (Hindle et al., 1993).

The method in resolving ambiguities that measures the strengths of association between two syntactic objects does not work well for some constructions, however. Consider the following examples:

(4) corn and *peanut butter*
(5) *put* the dress *on the rack*
(6) *Rangoon's* north *outskirts*

In (4), for instance, the strength of association for the two words *corn* and *peanut* may be greater than that of *corn* and *butter*, but the correct syntactic structure should be [*corn and* [*peanut butter*]].

Instead of measuring the strengths of association between two syntactic objects, we in this paper propose the use of co-occurrences among three syntactic objects involved in ambiguous constructions as an indicator of lexical preference. We then devise a way of measuring the strengths of association among the three words and finally apply the method to a disambiguation experiment in Japanese noun phrases with particle 'の' (*no*).

2 Dependency Relations and Class-based Estimation Method

2.1 Three Word Dependency Relations

A three word dependency relation (TWDR) is a syntactic relation that holds among any three words in a construction. It may be defined as:

<three word dependency relation> ::=
[W[<two word dependency relation>]] | [[<two word dependency relation>]W] |
[[W] <two word dependency relation>] | [<two word dependency relation>[W]]

<two word dependency relation> ::= [W[W]] | [[W]W]

Here W stands for any head word in a constituent and we understand that for any constituent d_i, $[d_1[d_2]]$ means that d_1 modifies d_2 and $[[d_1]d_2]$ means that d_2 modifies d_1.

A word may modify other word to its right or to its left in a construction. In a language like Japanese, however, a word always modifies another word to its right: we will have two possible dependency relations for the three word dependency construction. They are $[[w_1w_2]w_3]$ and $[w_1[w_2w_3]]$, where w_i stands for a head word in a constituent.

2.2 Estimation for the Strength of Lexical Association

There is a standard way to measure the strength of lexical association between any two words appearing in text. It is the mutual information (Church et al., 1990). The equations (1) and (2) show the ways of getting the mutual information between $\langle w_1 \rangle$ and $\langle w_2, w_3 \rangle$ for $[[w_1 w_2] w_3]$ and $[w_1 [w_2 w_3]]$.

$$I([[w_1 w_2] w_3]) = log_2 \left(\frac{N * f([[w_1 w_2] w_3])}{f(w_1) f(w_2 w_3)} \right) \tag{1}$$

$$I([w_1 [w_2 w_3]]) = log_2 \left(\frac{N * f([w_1 [w_2 w_3]])}{f(w_1) f(w_2 w_3)} \right) \tag{2}$$

here f is the frequency of co-occurrences and N the size of the training corpus.

A difficulty encountered in the calculation of the mutual information is the sparse data problem. How do we get reliable statistical results when no or only few word co-occurrences are observed? Among many (e.g., Alves, 1996; Jelinek et al., 1985), an answer to the question is to use word classes that contain the words in question and calculate the mutual information based on the word class co-occurrences. To deal with the sparse data problem, the equations (1) and (2) may be replaced by (3) and (4) :

$$I([[w_1, w_2] w_3]) \approx log_2 \left(\frac{N * f([[C_1, C_2] C_3])}{f(C_1) f([C_2, C_3])} \right) \tag{3}$$

$$I([w_1 [w_2, w_3]]) \approx log_2 \left(\frac{N * f([C_1 [C_2, C_3]])}{f(C_1) f([C_2, C_3])} \right) \tag{4}$$

Here C_i stands for a word class that includes the word w_i.

This way of estimating mutual information with word classes risks the problem of over-generalization, i.e., we may use a word class that is too general for the word in question. Here comes another problem: how do we choose the best class for the word concerned? The solution we offer for this problem is to choose word classes from a taxonomy using t-scores[1] as a measure of reliability.

For a class co-occurrence $\langle C_1; C_2, C_3 \rangle$, the t-scores for $[[C_1 C_2] C_3]$ and $[C_1 [C_2, C_3]]$ may be approximated by:

$$t([[C_1 C_2] C_3]) \approx \frac{f([[C_1 C_2] C_3]) - \frac{1}{N} f(C_1) * f(C_2, C_3)}{\sqrt{f([[C_1 C_2] C_3])}} \tag{5}$$

$$t([C_1 [C_2 C_3]]) \approx \frac{f([C_1 [C_2 C_3]]) - \frac{1}{N} f(C_1) * f(C_2, C_3)}{\sqrt{f([C_1 [C_2 C_3]])}} \tag{6}$$

t becomes very small or negative when $f([[C_1 C_2] C_3])$ or $f([C_1 [C_2 C_3]])$ is zero or low and it becomes bigger as the frequency becomes higher. t becomes also low when the classes in the co-occurrence contain too many words not relevant to the estimation.

The class-based estimation of mutual information using t-scores can be done in the following way:

(a) Set a threshold for t.

(b) Search for the word in question the lowest class co-occurrence in the taxonomy for which the t is above the threshold.

(c) Choose the most probable dependency structure using (3) and (4).

[1]The t-score (Church et al., 1993) is a standard measure of the likelihood that an occurrence can be attributed to chance.

3 An Application to Japanese Noun Phrases

The method of estimating the strengths of word association may be effectively used for resolving syntactic ambiguities. We would like to show an example in a disambiguation study that deals with a construction of Japanese noun phrases with particle *no*.

A function of particle *no* (の) in Japanese is in a sense similar to that of the preposition *of* in English. It builds up a noun phrase: *Tokyo no kita* (north of Tokyo). When more than two nouns are involved, a noun phrase with *no*'s becomes structurally ambiguous.

Theoretically a noun phrase with *no*'s can be infinite in its length. But the maximum number of *no* in a noun phrase with *no*'s is 4 in EDR Japanese Corpus[2].

Let us take an example and try to find the correct structure for three word *no*-constructions that can be either $[[w_1\ no\ w_2]\ no\ w_3]$ or $[w_1\ no\ [w_2\ no\ w_3]]$. Consider:

ラングーンの北のはずれ (Rangoon's north outskirts)

We first calculate the strength of association for [ラングーンの [北のはずれ]]. To do so, we replace the words in co-occurrence with their upper classes[3] according to the algorithm described in section 2 and obtain a co-occurrence with t-score just above the threshold (we set the threshold value to be .70 empirically).

Table 1 shows the conceptual hierarchy of *Rangoon*, *north* , and *outskirt*, and various frequencies of the concepts involved. Here, the numbers in columns $f(C_1)$, $f(C_2)$ and $f(C_3)$ are the occurrences in the EDR corpus for the classes in the conceptual hierarchy of the words *Rangoon*, *north* and *outskirts* respectively.

<table>
<tr><th colspan="3">co-occurrence</th><th>$f(C_1)$</th><th>$f(C_2)$</th><th>$f(C_3)$</th><th>$f(C_2C_3)$</th><th>$f([C_1[C_2C_3]])$</th><th>t</th></tr>
<tr><td>region</td><td>direction</td><td rowspan="2">location</td><td>181</td><td>42</td><td>513</td><td>10</td><td>1</td><td>.84</td></tr>
<tr><td rowspan="4">urban area</td><td rowspan="2">cardinal points</td><td>28</td><td>33</td><td>513</td><td>10</td><td>0</td><td>0</td></tr>
<tr><td rowspan="2">extreme</td><td>28</td><td>33</td><td>10</td><td>10</td><td>0</td><td>0</td></tr>
<tr><td rowspan="3">north</td><td>28</td><td>9</td><td>10</td><td>6</td><td>0</td><td>0</td></tr>
<tr><td rowspan="2">outskirt</td><td>28</td><td>9</td><td>2</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td>Rangoon</td><td>0</td><td>9</td><td>2</td><td>0</td><td>0</td><td>0</td></tr>
</table>

Table 1: Frequencies and t-score

In the case for[ラングーンの [北のはずれ]], we use the classes (region *no* direction *no* location) to estimate the mutual information. We repeat the same process for [[ラングーンの北] のはずれ] ([[Rangoon's north] outskirts]). The mutual information we get through this process is 2.58 for [ラングーンの [北のはずれ]] and 2.19 for [[ラングーンの北] のはずれ]. We are thus to select the structure [ラングーンの [北のはずれ]] as the most probable one.

4 Results and Evaluation

We have tested 429 *no*-constructions in EDR Corpus each of which contains two *no*'s. This number is somewhat arbitrary but all of the constructions are syntactically annotated. The training data, not including the test data, are 11,224 *no*-constructions that are all annotated.

Table 2 shows the disambiguation results (success rates) for the test data. Beside our method, we have tried to resolve ambiguities using various other methods including a human experiment. Here, the first result is the one obtained from attaching the modifier to the nearest modificant (head) (Kimball, 1973). The second one shows the performance by estimating the strength of association in two word dependency relations $\langle w_1, w_2\rangle$ and $\langle w_1, w_3\rangle$ (Alves, 1996; Kobayasi et al., 1994). The third one shows the performance of our method. And the last one is the performance by two native speakers who were just presented the *no*-constructions without surrounding contexts: we limited the test data to 150 among the 429 because it is too laborious to go through all of them.

[2] EDR Japanese Corpus is provided by Japan Electronic Dictionary Research Institute. It contains 220,000 parsed sentences with syntactic and semantic annotations.

[3] For this purpose, we use the Conceptual Dictionary provided by EDR Institute. It contains a hierarchical structure of all the words in the EDR corpus. The number of concepts in Conceptual Dictionary is about 400,000.

Method	Success
(1)Closest Attachment	73.2% (314/429)
(2)Two Word Dependency	72.7% (312/429)
(3)**Three word dependency**	77.6% (333/429)
(4)Human	80.0% (120/150)

Table 2: Experimental Results

The lower bound and the upper bound for the performance of our method seem respectively to be 73.2% by the simple heuristic of closest attachment and 80.0% by human beings (4). In the first case, it always attaches a word to its successor in a phrase. Nevertheless, it performed better than the method of using two word dependency relation. This result may be language specific because Japanese has a stronger tendency that a word tends to modify its adjacent word to its right. However, the performance of our method is much better than the first one. We attribute it to the estimation procedure that seems to capture the syntactic and semantic constraints better than other methods.

Although it is not a methodological deficiency, our method suffers apparently from the size of the training data (11,224): we found that about 37% of the test data suffer from data sparseness. When we eliminated these data and tested our method, we have acquired a success rate of 79.3%. This fact tells us that a larger training data is of help in improving the overall success rate.

5 Concluding Remarks

We in this paper presented a syntactic disambiguation method using lexical preferences estimated from three word dependency relations. It is proven by an experiment that the method worked well for determining the correct syntactic structure in ambiguous constructions.

We contend that the use of three word dependency relation in which we impose a finer constraint on word associations in predicting the correct structural attachment worked favorably in the disambiguation process. The t-score seems to have worked well also to selecting reliable classes from a taxonomical hierarchy. To our knowledge, the t-score, a standard measure of the likelihood that an occurrence can be attributed to chance, has not been used before as a measure of reliability in selecting class co-occurrences. We believe that the experimental result indicates an applicability of the method to resolving syntactic ambiguities in such constructions as nominal compounds, prepositional phrase attachments, and coordinate structures.

References

[1] Alves, E. (1996), "The Selection of the Most Probable Dependency Structure in Japanese Using Mutual Information." *Proceedings of the 34th Annual Meeting of ACL*, 372-374.

[2] Church, K., Mercer, R. (1993), "Introduction to the Special Issue on Computational Linguistics Using Large Corpora." *Computational Linguistics*, vol.19, no.1, 1-24.

[3] Church, K., Hanks, P. (1990), "Word Association Norms, Mutual Information and Lexicography." *Computational Linguistics*, vol.16, no.1, 22-29.

[4] Hindle, D., and Rooth, M. (1993), "Structural Ambiguity and Lexical Relations." *Computational Linguistics*, vol.19, no.1, 103-120.

[5] Jelinek, F., Mercer, R. (1985), "Probability Distribution Estimation from Sparse Data." *IBM Technical Disclosure Bulletin* vol.28, 2591-2594.

[6] Kimball, J. (1973), "Seven Principles of Surface Structure Parsing in Natural Language." *Cognition*, vol.2, 15-47.

[7] Kobayasi, Y., Tokunaga, T., Tanaka, H. (1996), "Analysis of Japanese Compound Noun Using Collocational Information." *Journal of Natural Language Processing*, vol.3, no.1, 29-44.

[8] Resnik, P. (1993), "Selection and Information: A Class-Based Approach to Lexical Relationships." PhD. thesis, University of Pennsylvania.