Structural Disambiguation Based on Reliable Estimation of Strength of Association

Haodong Wu Eduardo de Paiva Alves
Teiji Furugori
Department of Computer Science
University of Electro-Communications
1-5-1, Chofugaoka, Chofu, Tokyo 1828585, JAPAN
{wu,ealves,furugori}@phaeton.cs.uec.ac.jp

Abstract

This paper proposes a new class-based method to estimate the strength of association in word co-occurrence for the purpose of structural disambiguation. To deal with sparseness of data, we use a conceptual dictionary as the source for acquiring upper classes of the words related in the co-occurrence, and then use t-scores to determine a pair of classes to be employed for calculating the strength of association. We have applied our method to determining dependency relations in Japanese and prepositional phrase attachments in English. The experimental results show that the method is sound, effective and useful in resolving structural ambiguities.

1 Introduction

The strength of association between words provides lexical preferences for ambiguity resolution. It is usually estimated from statistics on word co-occurrences in large corpora (Hindle and Rooth, 1993). A problem with this approach is how to estimate the probability of word co-occurrences that are not observed in the training corpus. There are two main approaches to estimate the probability: smoothing methods (e.g., Church and Gale, 1991; Jelinek and Mercer, 1985; Katz, 1987) and class-based methods (e.g., Brown *et al.*, 1992; Pereira and Tishby, 1992; Resnik, 1992; Yarowsky, 1992).

Smoothing methods estimate the probability of the unobserved co-occurrences by using frequencies of the individual words. For example, when *eat* and *bread* do not co-occur, the probability of $\langle eat, bread\rangle$ would be estimated by using the frequency of $\langle eat\rangle$ and $\langle bread\rangle$. A problem with this approach is that it pays no attention to the distributional characteristics of the individual words in question. Using this method, the probability of $\langle eat, bread\rangle$ and $\langle eat, cars\rangle$ would become the same when *bread* and *cars* have the same frequency. It is unacceptable from the linguistic point of view.

Class-based methods, on the other hand, estimate the probabilities by associating a class with each word and collecting statistics on word class co-occurrences. For instance, instead of calculating the probability of $\langle eat, bread\rangle$ directly, these methods associate *eat* with the class $[ingest]$ and *bread* with the class $[food]$ and collect statistics on the classes $[ingest]$ and $[food]$. The accuracy of the estimation depends on the choice of classes, however. Some class-based methods (e.g., Yarowsky, 1992) associate each word with a single class without considering the other words in the co-occurrence. However, a word may need to be replaced by different class depending on the co-occurrence. Some classes may not have enough occurrences to allow a reliable estimation, while other classes may be too general and include too many words not relevant to the estimation. An alternative is to obtain various classes associated in a taxonomy with the words in question and select the classes according to a certain criteria.

There are a number of ways to select the classes used in the estimation. Weischedel *et al.* (1993) chose the lowest classes in a taxonomy for which the association for the co-occurrence can be estimated. This approach may result in unreliable estimates, since some of the class co-occurrences used may be attributed to chance. Resnik (1993) selected all pairs of classes corresponding to the head of a prepositional phrase and weighted them to bias the computation of the association in favor of higher-frequency co-occurrences which he considered "more reliable." Contrary to this assumption, high frequency co-occurrences are unreliable when the probability that the co-occurrence may be attributed to chance is high.

In this paper we propose a class-based method that selects the lowest classes in a taxonomy for which the co-occurrence confidence is above a threshold. We subsequently apply the method to solving structural ambiguities in Japanese dependency structures and English prepositional phrase attachments.

2 Class-based Estimation of Strength of Association

The strength of association (SA) may be measured using the frequencies of word co-occurrences in large corpora. For instance, Church and Hanks (1990) calculated SA in terms of mutual information between two words w_1 and w_2:

$$I(w_1, w_2) = log_2 \frac{N * f(w_1, w_2)}{f(w_1) f(w_2)} \qquad (1)$$

here N is the size of the corpus used in the estimation, $f(w_1, w_2)$ is the frequency of the co-occurrence, $f(w_1)$ and $f(w_2)$ that of each word.

When no co-occurrence is observed, SA may be estimated using the frequencies of word classes that contain the words in question. The mutual information in this case is estimated by:

$$I(C_1, C_2) = log_2 \frac{N * f(C_1, C_2)}{f(C_1) f(C_2)} \qquad (2)$$

here C_1 and C_2 are the word classes that respectively contain w_1 and w_2, $f(C_1)$ and $f(C_2)$ the numbers of occurrences of all the words included in the word classes C_1 and C_2, and $f(C_1, C_2)$ is the number of co-occurrences of the word classes C_1 and C_2.

Normally, the estimation using word classes needs to select classes, from a taxonomy, for which co-occurrences are *significant*. We use t-scores for this purpose[1].

For a class co-occurrence $\langle C_1, C_2 \rangle$, the t-score may be approximated by:

$$t \approx \frac{f(C_1, C_2) - \frac{1}{N} f(C_1) f(C_2)}{\sqrt{f(C_1, C_2)}} \qquad (3)$$

We use the lowest class co-occurrence for which the confidence measured with t-scores is above a threshold [2]. Given a co-occurrence containing the word w, our method selects a class for w in the following way:

```
Step 1: Obtain the classes $C^1, C^2, ..., C^n$ associated with $w$ in a taxonomy.
Step 2: Set $i$ to 0.
Step 3: Set $i$ to $i + 1$.
Step 4: Compute t using formula (3).
Step 5: If $t < threshold$.
            If $i \neq n$ goto step 3.
                Otherwise exit.
Step 6: Select the class $C^i$ to replace $w$.
```

Let us see what this means with an example. Suppose we try to estimate SA for $\langle produce, telephone \rangle$[3]. See Table 1. Here $f(v)$, $f(n)$ and $f(vn)$ are the frequencies for the verb *produce*, classes for the noun *telephone*, and co-occurrences between the verb and the classes for *telephone*, respectively; and t is the t-score[4].

[1] The t-score (Church and Mercer, 1993) compares the hypothesis that a co-occurrence is *significant* against the *null hypothesis* that the co-occurrence can be attributed to chance.

[2] The default threshold for t-score is 1.28 which corresponds to a confidence level of 90%. t-scores are often inflated due to certain violations of assumptions.

[3] The data was obtained from 68,623 verb-noun pairs in EDR Corpus (EDR, 1993).

[4] In our theory, we are to use each pair of $\langle C^i, C^j \rangle$, where i=1,2,...m, j=1,2,...,n, to calculate strengths of lexical associations. But our experiments show that upper classes of a verb are very unreliable to be used to measure the strengths. The reason may be that, unlike nouns, the verbs would not have a "neat" hierarchy or that the upper classes of a verb become too general as they contain too many concepts underneath them. Because of this observation, we use, for the classes of a

verb	classes for telephone	f(v)	f(n)	f(vn)	t-score
produce	concrete thing	671	18926	100	-4.6
produce	inanimate object	671	5593	69	0.83
produce	implement/tool	671	2138	35	1.91
produce	machine	671	664	19	2.86
produce	communication machine	671	83	1	0.25
produce	telephone	671	24	0	-

Table 1 Estimation of ⟨*produce telephone*⟩

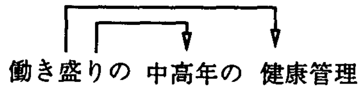
The lowest class co-occurrence ⟨*produce, communication machine*⟩ has a low t-score and produces a bad estimation. The most frequent co-occurrence ⟨*produce*, *concrete thing*⟩ has a low t-score also reflecting the fact that it may be attributed to chance. The t-scores for ⟨*produce, machine*⟩ and ⟨*produce, implement/tool*⟩ are high and show that these co-occurrences are significant. Among them, our method selects the lowest class co-occurrence for which the t-score is above the threshold: ⟨*produce, machine*⟩.

3 Disambiguation Using Class-Based Estimation

We now apply our method to estimate SA for two different types of syntactic constructions and use the results in resolving structural ambiguities.

3.1 Disambiguation of Dependency Relations in Japanese

Identifying the dependency structure of a Japanese sentence is a difficult problem since the language allows relatively free word orders. A typical dependency relation in Japanese appears in the form of modifier-particle-modificand triplets. When a modifier is followed by a number of possible modificands, there arise situations in which syntactic rules may be unable to determine the dependency relation or the modifier-modificand relation. For instance, in



'働き盛り'(vigorous) may modify either '中高年' (middle aged) or '健康管理' (health care). But which one is the modificand of '働き盛り'? We solve the ambiguity comparing the strength of association for the two or more possible dependency relations.

Calculation of Strength of Association We calculate the Strength of Association (SA) score for $modifier - particle - modificand$ by:

$$SA(m_f; p_{art}, m_c) = log_2\left(\frac{N * f(C_{mfier}, p_{art} m_c)}{f(C_{mfier}) f(p_{art} m_c)}\right) \quad (4)$$

where C_{mfier} stands for the classes that include the modifier word, p_{art} is the particle following the modifier, m_c the content word in the modificand phrase, and f the frequency.

Let us see the process of obtaining SA score in an example ⟨ 教授 - が - 働く ⟩ (literally: professor - subject-marker - work). To calculate the frequencies for the classes associated with '教授', we obtain from the Co-occurrence Dictionary (COD)[5] the number of occurrences for ⟨w- が -

verb, the verb itself or, when it does not give us a good result, only the lowest class of the verb in calculating the strength of association (SA). Thus, for an example, the verb eat has a sequence of *eat* → *ingest* → *put something into body* →.... → *event* → *concept* in the class hierarchy, but we use only *eat* and *ingest* for the verb eat when calculating SA for ⟨*eat, apple*⟩.

[5] COD and CD are provided by Japan Electronic Dictionary Research Institute (EDR, 1993). COD contains the frequencies of individual words and of the modifier-

働く〉, where w can be any modifier. We then obtain from the Concept Dictionary (CD)[6] the classes that include ‘ 教授 ’ and then sum up all the occurrences of words included in the classes. The relevant portion of CD for ‘ 教授 ’ in 〈 教授 - が - 働く 〉 is shown in Figure 1. The numbers in parenthesis here indicate the summed-up frequencies.

We then calculate the t-score between ‘ が - 働く ’ and all the classes that include ‘ 教授 ’. See Table 2.

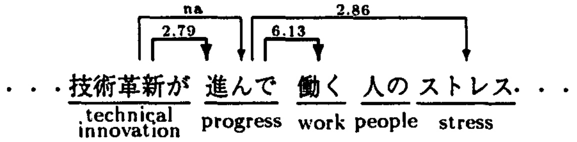
Classes for the modifier 教授	t-score	particle-modificand
人間と似た振舞をする主体	4.57	が働く
人間	5.14	が働く
役割で捉えた人間	1.74	が働く
身分で捉えた人間	0.74	が働く

Table 2 t-scores for 〈 教授 - が - 働く 〉

The t-score for the co-occurrence of the modifier and particle-modificand pair, ‘ 教授 ’ and ‘ が - 働く ’, is higher than the threshold when ‘ 教授 ’ is replaced with [役割で捉えた人間]. Using (4), the strength of association for the co-occurrence of 〈 教授 - が - 働く 〉 is calculated from the SA between the class [役割で捉えた人間] and ‘ が - 働く.’

When the word in question has more than one sense, we estimate SA corresponding to each sense and choose the one that results in the highest SA score. For instance, we estimate SA between ‘ 教授 ’ and the various senses of ‘ 働く ’, and choose the highest value: in this case the one corresponding to the sense ’to be employed.’

Determination of Most Strongly Associated Structure After calculating SA for each possible construction, we choose the construction with highest SA score as the most probable structure. See the following example:



Here, the arrows show possible dependency relations, the numbers on the arrows the estimated SA, and the thick arrows the dependency with highest mutual information that means the most probable dependency relation. In the example, ‘ 技術革新が ’ modifies ‘ 進んで ’ and ‘ 働く ’ modifies ‘ 人 ’. The estimated mutual information for 〈 技術革新が, 進んで 〉 is 2.79 and that for 〈 働く, 人 〉 is 6.13. Thus, we choose ‘ 進んで ’ as the modificand for ‘ 技術革新が ’ and ‘ 人 ’ as that for ‘ 働く ’.

In the example shown in Figure 2, our method selects the most likely modifier-modificand relation.

Experiment Disambiguation of dependency relations was done using 75 ambiguous constructions from Fukumoto (1992). Solving the ambiguity in the constructions involves choosing among two or more modifier-particle-modificand relations. The training data consists of all 568,000 modifier-particle-modificand triplets in COD.

Evaluation We evaluated the performance of our method comparing its results with those of other methods using the same test and training data. Table 3 shows the various results (success rates). Here, (1) indicates the performance obtained using the principle of Closest Attachment (Kimball, 1973); (2) shows the performance obtained using the lowest observed class co-occurrence (Weischedel *et al.*, 1993); (3) is the result from the maximum mutual information over all pairs of classes corresponding to the words in the co-occurrence (Resnik, 1993; Alves, 1996); and (4) shows the performance of our method[7].

particle-modificand triplets in a corpus that includes 220,000 parsed Japanese sentences.

[6] CD provides a hierarchical structure of concepts corresponding to all the words in COD. The number of concepts in CD is about 400,000.

[7] The precision is for the 1.28 default threshold. The precision was 81.2% and 84.1% when we set the threshold to .84 and .95. In all these cases the coverage was 92.0%.

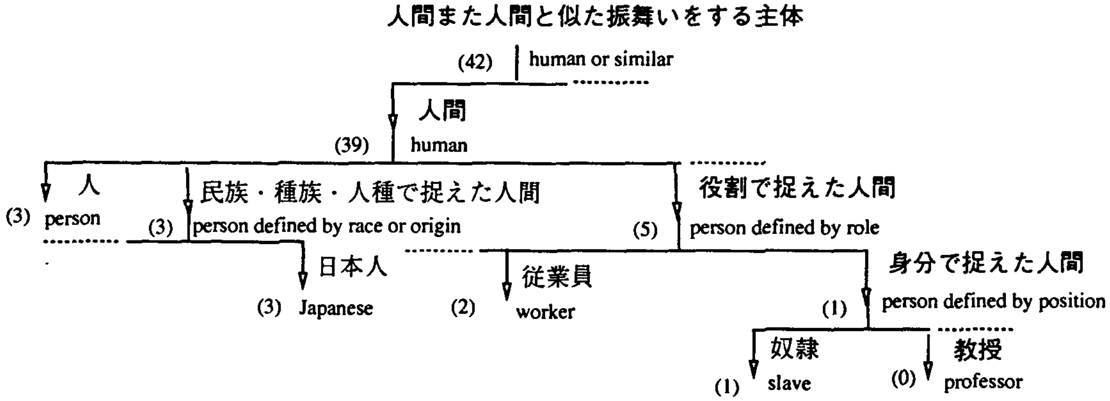


Figure 1 An Extract of CD

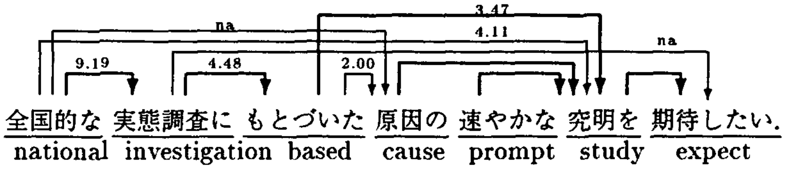


Figure 2 An example of parsing a Japanese sentence

method	precision
(1) closest attachment	70.6%
(2) lowest classes	81.2%
(3) maximum MI	82.6%
(4) our method	87.0%

Table 3 Results for determining dependency relations

Closest attachment (1) has a low performance since it fails to take into consideration the identity of the words involved in the decision. Selecting the lowest classes (2) often produces unreliable estimates and wrong decisions due to data sparseness. Selecting the classes with highest mutual information (3) results in overgeneralization that may lead to incorrect attachments. Our method avoids both estimating from unreliable classes and overgeneralization and results in better estimates and a better performance.

A qualitative analysis of our results shows two causes of errors, however. Some errors occurred when there were not enough occurrences of the particle-modificand pattern to estimate any of the strength of association necessary for resolving ambiguity. Other errors occurred when the decision could not be made without surrounding context.

3.2 Prepositional Phrase Attachment in English

Prepositional phrase (PP) attachment is a paradigm case of syntactic ambiguity. The most probable attachment may be chosen comparing the SA between the PP and the various attachment elements. Here SA is measured by:

$$SA(v\_attach|v,p,n_2) = log_2\left(\frac{N * f(C_v,p,C_{n_2})}{f(C_v)f(p,C_{n_2})}\right) \tag{5}$$

$$SA(n\_attach|n_1,p,n_2) = log_2\left(\frac{N * f(C_{n_1},p,C_{n_2})}{f(C_{n_1})f(p,C_{n_2})}\right) \tag{6}$$

where C_w stands for the class that includes the word w and f is the frequency in a training data containing *verb-noun1-preposition-noun2* constructions.

Our method selects from a taxonomy the classes to be used to calculate the SA score and then chooses the attachment with highest SA score as the most probable.

Experiment We performed a PP attachment experiment on the data that consists of all the 21,046 semantically annotated *verb-noun-preposition-noun* constructions found in EDR English Corpus. We set aside 500 constructions for test and used the remaining 20,546 as training data. We first performed the experiment using various values for the threshold. Table 4 shows the results. The first line here shows the default which corresponds to the most likely attachment for each preposition. For instance, the preposition *of* is attached to the noun, reflecting the fact that PP's led by *of* are mostly attached to nouns in the training data. The 'confidence' values correspond to a binomial distribution and are given only as a reference[8].

confidence	t	coverage	precision	success
-	-	100%	68.0%	68.0%
50%	.00	82%	82.2%	79.4%
70%	.52	75%	87.3%	83.4%
80%	.84	65%	88.6%	84.2%
85%	.95	57%	89.6%	84.8%
90%	1.28	50%	91.3%	85.6%

Table 4 Results for PP attachment with various thresholds for t-score

The precision grows with t-scores, while coverage decreases. In order to improve coverage, when the method cannot find a class co-occurrence for which the t-score is above the threshold, we recursively tried to find a co-occurrence using the threshold immediately smaller (see Table 4). When the method could not find co-occurrences with t-score above the smallest threshold, the default was used. The overall success rates are shown in "success" column in Table 4.

Evaluation We evaluated the performance of our method comparing its results with those of other methods with the same test and training data. The results are given in Table 5. Here, (5) shows the performance of two native speakers who were just presented quadruples of four head words without surrounding contexts.

Method	Success Rate
(1)closest Attachment	59.6%
(2)lowest classes	80.2%
(3)maximum MI	79.0%
(4)our method	85.6%
(5)human (head words only)	87.0%

Table 5 Comparison with other methods

The lower bound and the upper bound on the performance of our method seem to be 59.6% scored by the simple heuristic of closest attachment (1) and 87.0% by human beings (4). Obviously, the success rate of closest attachment (1) is low as it always attaches a word to the noun without considering the words in question. The unanticipated low success rate of human judges is partly due to the fact that sometimes constructions were inherently ambiguous so that their choices differed from the annotation in the corpus.

Our method (4) performed better than the lowest classes method (2) and maximum MI method (3). It owes mainly to the fact that our method makes the estimation from class co-occurrences that are more reliable.

4 Concluding Remarks

We proposed a class-based method that selects classes to be used to estimate the strength of association for word co-occurrences. The classes selected by our method can be used to estimate various types of strength of association in different applications. The method differs from other class-based methods in that it allows identification of a reliable and specific class for each co-occurrence in consideration and can deal with date sparseness problem more efficiently. It

[8] As another way of reducing the sparse data problem, we clustered prepositions using the method described in Wu and Furugori (1996). Prepositions like synonyms and antonyms are clustered into groups and replaced by a representative preposition (e.g., *till* and *pending* are replaced by *until*; *amongst*, *amid* and *amidst* are replaced by *among*.).

overcame the shortcomings from other methods: overgeneralization and employment of unreliable class co-occurrences.

We applied our method to two structural disambiguation experiments. In both experiments the performance is significantly better than those of others.

References

[1] Alves, E. 1996. "The Selection of the Most Probable Dependency Structure in Japanese Using Mutual Information." In *Proc. of the 34th ACL*, pages 372-374.

[2] Brown, P., Della Pietra, V. and Mercer, R. (1992). "Word Sense Disambiguation Using Statistical Methods." *Proceedings of the 30th ACL*, pages 264-270.

[3] Church, K., and Mercer, R. 1993. "Introduction to the Special Issue on Computational Linguistics Using Large Corpora." *Computational Linguistics*, 19(1):1-24.

[4] Church, K., and Hanks, P. 1990. "Word Association Norms, Mutual Information and Lexicography." *Computational Linguistics*, 16(1):22-29.

[5] Church, K., and Gale, W. 1991. "A Comparison of the Enhanced Good-Turing and Deleted Estimation Methods for Estimating Probabilities of English Bigrams." *Computer Speech and Language*, 5:19-54.

[6] Fukumoto, F., Sano, H., Saitoh, Y. and Fukumoto J. 1992. "A Framework for Dependency Grammar Based on the Word's Modifiability Level - Restricted Dependency Grammar." *Trans. IPS Japan*, 33(10):1211-1223 (in Japanese).

[7] Hindle, D., and Rooth, M. 1993. "Structural Ambiguity and Lexical Relations." *Computational Linguistics*, 19(1):103-120.

[8] Japan Electronic Dictionary Research Institute, Ltd. 1993. *EDR Electronic Dictionary Specifications Guide* (in Japanese).

[9] Jelinek, F., and Mercer, R. 1985. "Probability Distribution Estimation from Sparse Data." *IBM Technical Disclosure Bulletin*, 28:2591-2594.

[10] Katz, S. 1987. "Estimation of Probabilities from Sparse Data for Language Model Component of a Speech Recognizer." *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-35(3):400-401.

[11] Kimball, J. 1973. "Seven Principles of Surface Structure Parsing in Natural Language." *Cognition*, 2:15-47.

[12] Pereira, F. and Tishby, N. 1992. "Distributional Similarity, Phrase Transitions and Hierarchical Clustering." In *Proc. of the 30th ACL*, pages 183-190.

[13] Resnik, P. 1992. "Wordnet and Distributional Analysis: A Class-Based Approach to Lexical Discovery." *AAAI Workshop on Statistically-based Natural Language Processing Techniques*, pages 56-64.

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

[15] Weischedel, R., Meteer, M., Schwartz, R., Ramshaw, L., and Palmucci, J. 1993. "Coping with Ambiguity and Unknown Words Through Probabilistic Models." *Computational Linguistics*, 19(2):359-382.

[16] Wu, H. and Furugori, T. 1996. "A Hybrid Disambiguation Model for Prepositional Phrase Attachment." *Literary and Linguistic Computing*. 11(4):187-192.

[17] Yarowsky, D. 1992. "Word Sense Disambiguation using Statistical Models of Roget's Categories Trained on Large Corpora." *Proceedings of COLING-92*, pages 454-460.