自 然 言 語 処 理　135－21
（2000．1．28）

Clustering Similar Words for Curing Data Sparseness

Haodong Wu, Takahiro Azuma, Hisao Matumaru　　**Teiji Furugori**

Department of Language and Culture
Dokkyo University

Department of Computer Science
University of Electro-Communications

Abstract

The sparse data problem is a notorious problem in statistical linguistics. In this paper, we describe two techniques to reduce this problem. First, we propose a method that measures contextual similarity with the assumption that two words are similar if they have similar average mutual information. Then, we present a class-based approach to select class co-occurrences in replace of unobserved lexical co-occurrences. These classes are acquired from an IS-A taxonomy using t-score as a reliability measure. We then employ them for syntactic disambiguation tasks. The experiments that follow show effectiveness of our approach and their applicability for many applications in Natural Language Processing.

類似語のクラスタリングによるデータ希薄性の解消

呉 浩東　東　孝博　松丸壽雄
外国語学部言語文化学科
獨協大学

古郡廷治
情報工学科
電気通信大学

概　要

希薄なデータ問題は統計的な言語処理における難題の一つである。本稿は二つの手法を提案してこの問題を対処する。本稿はまず、相互情報量の類似性と単語間類似性を関連性から、文脈的語彙類似性の尺度を定義する。次に、t-score は信頼性をはかる尺度として、シソーラスから信頼性の高いクラスを選ぶ手法を提案する。この二つの手法を統語的な曖昧性解消に用いて実験を行い、その有効性を実証した。

1. Introduction

Statistical data on word co-occurrence relations play an important role in corpus-based approaches for natural language processing. The co-occurrence relations include co-occurrences within n-grams (consecutive sequence of n words), with in certain syntactic relation (i.e., adjective-noun, subjective-verb) or within a local context. Statistic data with these co-occurrences are widely used in information retrieval, semantic clustering or automatic thesaurus construction, machine translation, syntactic and lexical disambiguation tasks.

A major problem in the above applications is how to estimate the probabilities of word co-occurrences that were not observed or the observed co-occurrences are too low in the training corpus. This problem is known as data sparseness and occurs often (more than 25% cases) even using a very large corpus (Church & Mercer, 1993).

There are two major approaches in the literature for resolving the sparse-data problem. They

are smoothing and class-based methods. Smoothing methods estimate the probability of unobserved co-occurrences using frequency information (e.g., Good 1953, Katz 1987; Church & Gale 1991). Good (1953) used Turing's formula as an improved estimate over the maximum likelihood estimation (MLE) method. Katz (1987) proposed a so-called back-off model for smoothing unreliably estimated parameters with their correlated (n-1)-gram parameters. Church and Gale (1991) suggested a technique of using *frequency-based estimation* for smoothing estimates of unobserved co-occurrences, an estimate based solely on the frequencies of either the frequency of co-occurrence (the frequency of n-gram), or the frequency of the individual words which compose the co-occurrence.

Class-based techniques (e.g., Resnik 1993, 1999; Dagan et al. 1995) estimate unobserved co-occurrences using classes of "similar" words. This approach collects statistics not on individual words but on classes of similar words. It is based on an implicit assumption that words that behave in the same way on the surface have semantic similarities. Thus, the probability of a specific co-occurrence is determined using generalized parameters about the probability of a class co-occurrence. This approach suffers from the loss of too much information when the word co-occurrence patterns are generalized to class co-occurrence parameters, however.

We in this paper propose two kinds of method to reduce the sparse data problem: one is to cluster similar words from a large scale corpus; and another one is to use a toxonomy to acquire reliable class co-occurrences. We then use them to acquire lexical preferences for resolving English prepositional phrase attachments and coordinate construction ambiguities.

2. Clustering Words Using Contextual Similarity Estimated from Corpus

Any approach that relies on statistical data about lexical co-occurrence faces the sparse data problem, which makes it very difficult to estimate probability of certain lexical co-occurrence, and it in consequence limits the coverage of the estimation, or reduces the precision due to inaccurate estimates. To deal with this problem, we suggest a clustering approach that replaces the words in certain co-occurrence with their similar words in the context. The goal of the method is to estimate the likelihood of co-occurrences that were not observed in the corpus. The basic idea is to estimate the likelihood of a given unobserved co-occurrence using the estimated probabilities of other co-occurrences that were observed in the corpus, and contain words are "similar" to the words of the observed co-occurrence. To measure the similarity between two words we use mutual information (Church & Hanks, 1990) as a measure for word association. The similarity between two words w_1 and w_2 is defined by using their next word and their preceding word (the word not including a determiner), denoted as w. We assume w_1 and w_2 are similar if they have similar mutual information with any adjacent word w in a large corpus. Since co-occurrence pairs are directional, we get two measures according to the position of w. The *left context similarity* of w_1 and w_2 is defined as (1).

$$sim_L(w_1, w_2) = \frac{\sum_{w \in lecical} (\min(\Gamma(w, w_1) \cdot I(w, w_1), \Gamma(w, w_2) \cdot I(w, w_2))}{\sum_{w \in lexical} (\max(\Gamma(w, w_1) \cdot I(w, w_1), \Gamma(w, w_2) \cdot I(w, w_2))} \quad (1)$$

$\Gamma(w, w_i)$ is the weight in [0.5, 1] according to importance of the syntactic relation between two words (i.e., verb-noun, adjective-noun). It can avoid choosing words in collocations such as *computer science, information technology* as similar words. The *right context similarity* is defined equivalently. The general definition of similarity between w_1 and w_2 are defined as:

$$sim(w_1, w_2) = \frac{sim_L(w_1, w_2) + sim_R(w_1, w_2)}{2} \quad (2)$$

3. Reliable Word Co-occurrences Estimation from Taxonomy and Corpora

This section presents another method to acquire the word classes using semantic similarity measure with a taxonomy. *Nurse* and *hospital*, for example, would seem to be more similar than *nurse* and *teacher* in the context, but the latter pair are certainly more similar in semantic point of view as it contains more features in their information content.

The problem we are confronted with is how to acquire similar word classes to replace the words in certain co-occurrence relationship without loss of too much information or features the words carried. Here, we use a IS-A taxonomy to acquire an appropriate class for word w. Many words, however, are polysemous and even belong to multiple parts of speech. We classify nouns to their upper classes prefer to other kinds of words because nouns in taxonomies tend to be classified finer than others words. If word w are polysemous and have n senses in the taxonomy, let T be the set of concepts (corresponding to word senses) in an IS-A taxonomy, permitting multiple inheritance. To choose the most possible sense(s) of w, we check its local context and decide it using the method described in (Wu & Furugori, 1999). We use an algorithm which employs multiple clues as contextual similarities, parts of speech, word co-occurrences in local context, syntactic relations to select the most possible word senses.

Now, we consider the measure to justify if the frequency of a certain lexical co-occurrence in a corpus is too low to make a reliable estimation. We use t-score as a measure of reliability.[1] (Alves, Wu & Furugori, 1998; Alves & Furugori, 1999). For a certain class co-occurrence $\langle C1 \rangle$ and $\langle C2 \rangle$ the t-score may be approximated by:

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

here N is the size of the corpus in the estimation, $C1$ and $C2$ are the word classes that respectively contain the word $w1$ and the word $w2$, $f(C1)$ and $f(C2)$ are the numbers of occurrences of all the words include in the word classes, $C1$ and $C2$ and $f(C1, C2)$ is the numbers of co-occurrences of the word classes $C1$ and $C2$. We set a threshold[2] for t-scores and use the lowest class co-occurrence for which the confidence measured with t-score is above the threshold. Given a co-occurrence containing $w1$ and $w2$, Algorithm 1 selects two classes for the two words.

Algorithm 1: Selecting the appropriate classes from a taxonomy

Step 1: set $C1$ = w1, $C2$ = w2

Step 2: while t($C1$, $C2$) < threshold-t

if the number of members in $C1$ < threshold-w1 then

replace $C1$ with its upper class in the taxonomy

if the number of members in $C2$ threshold-w2

---

[1] The t-score (Church and Merser, 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.

replace C_1 with its upper class in the taxonomy
else if the number of members in $C_1 \geqq$ threshold-w1 then exit.
Step 3: choose C_1 and C_2.

Here, threshold-w1 and threshold-w2 is experimentally set to 100 (verb, adjective or adverb) or 1000 (noun) to avoid using too abstract classes. To improve the classification, we use the following processing to deal with some special words.

1. All 4-digit numbers are replaced with 'date'.
2. All verbs with an affix of *re-* or *un-* are reduced to their stems.
3. Nouns starting with a capital letter are replaced with 'name'.
4. Personal pronouns are replaced with 'person'.

Algorithm 1 is also applicable for estimating word classes from n-grams.

4. Applications to Structural Ambiguity Resolution

We now apply the methods mentioned above to acquire lexical preferences for two different syntactic constructions and use the results for structural ambiguity resolutions.

4.1 Prepositional Phrase Attachment Using Reliable Class Co-occurrences

Prepositional Phrase (PP) attachment is a paradigm case of syntactic ambiguity. Like other work, we consider that the following head words of: main verb (v), head noun (n1) ahead of the preposition (p), and head noun (n2) of the object of the preposition to determine PP attachments.

We use lexical association acquired from two annotated corpora: EDR English Corpus (EEC)[3] and Susanne Corpus(SC). Each sentence in the corpora is syntactically tagged. This means that a given PP is attached to a unique phrase.

We use the RA (Ratio of Association) score to estimate the likelihood of attachment for a certain (v, n1, p, n2). The RA score is defined in (4) as a value of counts of VP attachments divided by the total occurrences of (v, n1, p, n2) in the training data.

$$RA(v,n1,p,n2) = \frac{f(vp \mid v,n1,p,n2)}{f(v,n1,p,n2)} \approx \frac{f(vp \mid v,n1,p,n2)}{f(vp \mid v,n1,p,n2) + f(np \mid v,n1,p,n2)} \quad (4)$$

the symbol f denotes frequency of a particular quadruple in the training data. The RA score ranges from 0 to 1. We choose the VP attachment when RA(v, n1, p, n2) > 0.5. Otherwise we choose the NP attachment.

The quadruples in test data are seldom found in the training data due to the data sparseness (about 94.3% of the quadruples in the test data are not seen in the training data). We therefore turn to collecting triplets of (v, p, n2), (n1, p, n2), (v, n1, p)[4] as Collins and Brooks (1995) did to smooth the estimation, and compute the RA score using formula 3.

$$RA(v,n1,p,n2) = \frac{f(vp \mid v,p,n2) + f(vp \mid n1,p,n2) + f(vp \mid v,n1,p)}{f(v,p,n2) + f(n1,p,n2) + f(v,n1,p)} \quad (5)$$

---

[3] EEC, compiled by Japan Electronic Dictionary Research Institute, Ltd., contains 160,000 sentences with annotated morphological, syntactic, and semantic information.

[4] As Collins and Brooks (1995) points out, triplets not containing prepositions do not work well in predicting PP attachment.

We find that the RA score makes an undesirable estimation in two cases. It is when the co-occurrences are too low (e. g. only one case found in the training data) or the score is close to the boundary value of 0. 5. We introduce two thresholds to deal with this problem.

For (5), the condition is: $f_{triplet}(v,n1,p,n2) \geq 3$, and

$$|2 \times RA(v,n1,p,n2) - 1| \times \log_2(f_{triplet}(v,n1,p,n2)) < 0.5$$

where $f_{triplet}(v,n1,p,n2)$ is defined as:

$$f_{triplet}(v,n1,p,n2) = f(v,p,n2) + f(n1,p,n2) + f(v,n1,p)$$

For each sentence with ambiguous PP(s), the disambiguation model is put in algorithm 2:

Algorithm 2: Determining PP attachment

Set C_v = v, C_{n1} = n1, C_p = p, C_{n2} = n2, Set $RA(v,n1,p,n2)$ = -1;

While $n(C_v) < 100$ and $n(C_{n1}) < 1000$ and $n(C_{n2}) < 1000$[5]

define $f_{triplet}(C_v,C_{n1},C_p,C_{n2}) = f(C_v,C_p,C_{n2}) + f(C_{n1},C_p,C_{n2}) + f(C_v,C_{n1},C_p)$

if $f_{triplet}(C_v,C_{n1},C_p,C_{n2}) \geqq 3$, then

$$RA(v,n1,p,n2) = \frac{f(vp|C_v,C_p,C_{n2}) + f(vp|C_{n1},C_p,C_{n2}) + f(vp|C_v,C_{n1},C_p)}{f_{triplet}(C_v,C_{n1},C_p,C_{n2})}$$

else if $|2 \times RA(v,n1,p,n2) - 1| \times \log_2(f_{triplet}(C_v,C_{n1},C_p,C_{n2})) < 0.5$

then {$RA(v,n1,p,n2) = -1$, and replace C_v,C_{n1},C_{n2} using Algorithm 1 and

replace C_p using its similar prepositions[6]. }

if $RA(v,n1,p,n2) \geq 0$, then {

if $RA(v,n1,p,n2) \geq 0.5$, then choose NP attachment;

else choose VP attachment. exit.}

if f(p)>0, then {

if $f(vp|p)/f(p) < 0.5$, then choose NP attachment;

else choose VP attachment.}

else choose NP attachment.

---

[5] $n(C)$ is the number of the members in class C.

[6] Similar prepositions are clustered into groups and replaced by a representative preposition in which is described in Wu and Furugori (1996). For example, prepositions of among, amongst, amid, amidst are clustered into a group and the pressentative preposition is among.

3.2 Experiment and Evaluation

We now describe an experiment used for testing our method. We used test data that contains 3043 ambiguous PPs extracted randomly from a computer manual, a grammar book and the *Japan Times*. The overall success rate is 86.7%. We have made a comparison of the performance of our method with that of other work. Table 1 gives the results of the comparison. Here, (1) shows the result using a structural heuristic — Right Association with our test data. (2) shows the result of an example-based method by Sumita et al. (1994), where both the training data (3299 handcrafted examples) and the test data are taken in a small domain. (3) and (4-1) show the results of two statistics-based methods (Brill et al. 1994; Collins et al. 1995), where both use the WSJ corpus as their training data and the IBM data as their test data.

Of these methods, our method is somewhat similar to Collins and Brooks (1995) in using triplet combinations. To make a comparison of efficiency, we have tried their method both on our training data and test data, and got a success rate of 84.2%, a result 2.5% lower than that of ours. The success rate of our method is consistently better than that of other work by 3% to 4%. It owes mainly to the fact that our method makes the estimation from class co-occurrences that are more reliable. If we do not use the class co-occurrences in place of the word co-occurrences, the success rate is 79.3%, 7.4% lower than the result turned out by algorithm 2. Thresholds on t-score and the sizes of the classes are also essential to choose the reliable classes.

Table 1: Comparison with Other Work

Method	Number Tested	Success Rate
(1) Right Association	3043	67.1%
(2) Example-based [SFI 94]	131	85.7%
(3) Rule-based [BR 94]	3097	81.9%
(4-1) Backed-off [CB 95] (original)	3097	84.5%
(4-2) Backed-off [CB 95] (using our data)	3043	84.2%
(5) This method	3043	86.7%
average native speakers (4 head words)	300	88.3%

4.3 Resolving Coordination Ambiguity Using Contextual Similarity

Resolving ambiguity in coordinate construction (CC) is to determine the way of conjoining constitutents (words, phrases, or clauses) and/or to determine the scope of coordination, i.e., immediacy relations among the continents involved. For instance, in *sweet and sour pork*, the right immediacy relation is *((sweet and sour) pork)* rather than *(sweet and (sour pork))*.

We propose Algorithm 3 to disambiguating CCs in the form of n1+and+n2+n3 (Wu & Furugori, 1998).

Algorithm 3: Disambiguations for coordinations in the form of n1+and+n2+n3

(1) If all of the three nouns are capitalized, produce ((n1 and n2) n3).

(2) Otherwise,

(a) if n1 and n2 match in number and n1 and n3 do not, produce ((n1 and n2) n3);

(b) if n1 and n3 match in number and n1 and n2 do not, produce (n1 and (n2 n3)).

(3) Otherwise,

(a) if n1 is the antonym[7] of n2, produce ((n1 and n2) n3);

[7] The antonyms are acquired using WordNet (Miller, 1990).

(b) if n1 is the antonym of n3, produce (n1 and (n2 n3)).

(4) Otherwise,

(a) if sim(n1, n2)>=sim(n2, n3)[8], produce ((n1 and n2) n3).

(b) if sim(n1, n2)<sim(n2, n3), produce (n1 and (n2 n3)).

To evaluate the performance of the disambiguation algorithms, we randomly selected 300 coordinate structures of the form of n1+n2+and+n3 from on-line CNN news. Table 2 shows the result for disambiguating CCs of n1+and+n2+n3. The overall success rate of Algorithm 3 is 85.3%.

The algorithms have adopted a backed-off form to integrate different cues in the disambiguation process and the reliable cues with certainty are used first to achieve a better overall performance. As we can see from the results, the cues (1) (2) (3) in table 2 with high success rates have comparatively low recall rates; semantic similarity, on the other hand, have high recall rates with lower success rate.

Table 2: Experimental Results on Testing n1+and+n2+n3

Step	Recall	Number	Accuracy
(1) orthographic form	3.7%	11	100.0%
(2) antonym	8.3%	24	95.2%
(3) similarity in form	38.8%	104	91.4%
(4) semantic similarity	100.0%	161	78.8%
Total	100.0%	300	85.3%

Table 3 shows the results of the performance achieved by our method and those by others for the structure of n1+and+n2+n3. All these methods use the same data. Here, (1) shows the result obtained from attaching the modifier to the nearest head (Kimball, 1973), i.e., (n1 and (n2 n3)); (2) shows the result of the method proposed by Resnik (1993); (3) shows the performance of our method; and (4) is the performance of human judgment by three native speakers who were just presented the words of n1, coordinator, n2 and n3 without surrounding contexts.

Table 3: Comparison with Other Work for Determining n1+and+n2+n3

Method	Success Rate
(1) Closest attachment	65.3%
(2) Resnik's method	80.7%
(3) Our method	85.3%
(4) Average human	91.3%

5. Conclusions

We have presented two methods: clustering similar words from corpus and class-based estimation with a taxonomy to reduce the sparse data problem. Using the first method, we can measure any two words with contextual similarity. We do not produce word classes directly but use the measure for disambiguation task of coordinate constructions. The second method extracted reliable word classes from a taxonomy using t-score can avoid using low occurrence of certain lexical

[8] The training data are extracted from The British Corpus (http://thetis.bl.uk) which contains about 95,890,000 words.

co-occurrence and too general classes which loss too much of information. Two disambiguation experiments have proven our methods are effective, robust and useful in resolving different types of ambiguities. The performances of our algorithms for the disambiguation tasks are significantly better than those of other related work. The methods can also be used for selecting 'correct' word sense for machine translation, constructing dynamic thesaurus on finding similar words related the input key word(s) for information retrieval. Yet, since the estimation methods are applicable to any types of lexical relationship, their performances may be further improved by using larger corpora and taxonomies.

References

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

[2] Church, K. W. and Gale, W. A. 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.

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

[4] Collins, M. and Brooks, J. 1995. "Prepositional Phrase Attachment Through a Backed-off Model." In *ACL-95 Corpus-based Workshop*.

[5] Dagan, I. ; Marcus, S. ; and Markovitch S. 1995. "Contextual Word Similarity and Estimation from Sparse Data." *Computer Speech and Language*, 9:123-152.

[6] Alves, E., Wu, H., and Furugori, T. 1998. "A Method for Estimating Strength of Association and Its Application to Structural Disambiguation." *Journal of Natural Language Processing*, 5(3):53-65.

[7] Alves, E. and Furugori, T. 1999. "Three-Word Dependency Relations and Their Application to Structural Ambiguity Resolution." *Transactions of Information Society of Japan*, 40(1):343-350.

[8] Good, I. J. 1953. "The Population Frequencies of Species and the Estimation of Population Parameters." *Biometrika* 40:237-263.

[9] Japan Electronic Dictionary Research Institute, Ltd. 1993. *EDR Electronic Dictionary Specifications Guide*.

[10]Katz, S. M. 1987. "Estimation of Probabilities from Sparse Data for the Language Modeling for Speech Recognizer." *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35:400-401.

[11]Miller, G. 1990. "WordNet: An On-line Lexical Database." *International Journal of Lexicography*, 3(4). (Special Issue).

[12]Resnik, P. 1993. "Selection and Information: A Class-Based Approach to Lexical Relationships." Doctoral Dissertation, University of Pennsylvania, Philadelphia, P.A.

[13]Resnik, P. 1999. "Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language." Journal of Artificial Intelligence Research, 11:95-130.

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

[15]Wu, H. and Furugori, T. 1998. "A Hybrid Approach for Resolving Ambiguities in Coordinate Structures." Natural Language Processing, 5(4):1-16.