Redefining similarity in a thesaurus by using corpora

Hiroyuki Shinnou
Ibaraki University
Dept. of Systems Engineering
Nakanarusawa, 4-12-1
Hitachi, Ibaraki, 316, Japan
shinnou@lily.dse.ibaraki.ac.jp

1 Introduction

The aim of this paper is to automatically define the similarity between two nouns which are generally used in various domains. By these similarities, we can construct a large and general thesaurus.

In applications of natural language processing, it is necessary to appropriately measure the similarity between two nouns. The similarity is usually calculated from a thesaurus. Since a handmade thesaurus is not suitable for machine use, and expensive to compile, automatical construction of a thesaurus has been attempted using corpora (Hindle, 1990). However, the thesaurus constructed by such ways does not contain so many nouns, and these nouns are specified by the used corpus. In other words, we cannot construct the general thesaurus from only a corpus. This can be regarded as data sparseness problem that few nouns appear in the corpus.

To overcome data sparseness, methods to estimate the distribution of unseen cooccurrence from the distribution of similar words in the seen cooccurrence has been proposed. Brown et al. proposed a class-based n-gram model, which generalizes the n-gram model, to predict a word from previous words in a text (Brown et al., 1992). They tackled data sparseness by generalizing the word to the class which contains the word. Pereira et al. also basically used the above method, but they proposed a soft clustering scheme, in which membership of a word in a class is probabilistic (Pereira et al., 1993). Brown and Pereira provide the clustering algorithm assigning words to proper classes, based on their own models. Dagan et al. proposed a similarity-based model in which each word is generalized, not to its own specific class, but to a set of words which are most similar to it (Dagan et al., 1993). Using this model, they successfully predicted which unobserved cooccurrences were more likely than others, and estimated the probability of the cooccurrences (Dagan et al., 1994). However, because these schemes look for similar words in the corpus, the number of similarities which we can define is rather small in comparison with the number of similarities for pairs of the whole. The scheme to look for similar words in the corpus has already taken the influence of data sparseness.

In this paper, we propose a method distinct from the above methods, which use a handmade thesaurus to find similar words. The proposed method avoids data sparseness by estimating undefined similarities from the similarity in the thesaurus and similarities defined by the corpus. Thus, the obtained similarities are the same in number as the similarities in the thesaurus, and they reflect the particularity of the domain to which the used corpus belongs. The use of a thesaurus can obviously set up the similar word independent of the corpus, and has an advantage that some ambiguities in analyzing the corpus are solved.

We have experimented by using Bunrui-goi-hyou(Bunrui-goi-hyou, 1994), which is a kind of Japanese handmade thesaurus, and the corpus which consists of Japanese economic newspaper 5 years articles with about 7.85 M sentences. We evaluate the appropriateness of the obtained similarities.

2 Defining the similarity

We can easily judge the similarity of two nouns if they are very similar. However, the more different they are, the more difficult it is to define their similarity. Thus, we can trust that nouns in the class corresponding to the "leaf" of Bunrui-goi-hyou are similar to each another, and this is not affected by the domain. In this paper, we will refer to the class corresponding to the leaf of Bunrui-goi-hyou **the primitive class**. Therefore, the similarity we have to define is the similarity between these classes.

This method consists of 4 steps.

Step 1 Gather the cooccurrence data from the corpus.

Step 2 Generalize the noun in the cooccurrence data to the primitive class.

Step 3 Measure the similarity between two primitive classes by using the cooccurrence data obtained in step 2.

Step 4 Estimate undefined similarities.

We will describe each step in detail in following subsections.

2.1 Gathering cooccurrence data (step 1)

In order to carry out our method, it is necessary to first gather the cooccurrence data from the corpus.

If a noun (N), a postpostional particle (P), and a verb (V) appear in a sentence in this order, we pick out the cooccurrence data $[N, P, V]$. In this study, we gathered cooccurrence data only from the postpostional particle "wo", because "wo" is the most effective postpostional particle for classifying nouns.

As a corpus, we used five years of Japanese economic newspaper articles. The corpus has about 7.85 M sentences, and the average number of characters in one sentence was about 49. From the corpus, we gathered about 4.41 M bits of cooccurrence data (about 1.48 M types) whose postpositional particle was "wo". From them, we removed the cooccurrence data whose frequency was 1, or whose verb does not appear more than 20 times. In all, we obtained about 3.26 M bits of cooccurrence data, which consisted of about 0.36 M types. These cooccurrence data are used in the next step.

2.2 Generalizing the word to the class (step 2)

In step 2, we generalize the noun in cooccurrence data gathered in step 1 to the primitive class to which this noun belongs.

First, we should explain about Bunrui-goi-hyou. Bunrui-goi-hyou is a kind of thesaurus with a tree-like structure that has a maximum depth of level 6. Class IDs are assigned to each "leaf" of the "tree". Each noun has a class ID corresponding to the meaning of the noun. The class ID corresponds to the primitive class. Bunrui-goi-hyou has 3,582 primitive classes.

Because many nouns, such as compound nouns, are not in Bunrui-goi-hyou, we cannot always generalize all nouns to primitive classes, 86.0% of the nouns in cooccurrence data gathered in step 1 could be generalized to primitive classes.

In this generalization, the problem of polysemy arises. A noun has usually several primitive classes because of the polysemy. We solve some polysemy from the distribution of nouns in cooccurrence data which have the same verb. This cannot be discussed here for lack of space. We only report that the cooccurrence data gathered in step 1 contain 572,529 bits of polysemy which consisted of 27,918 types, and 472,273 bits of polysemy (18,534 types) were solved.

In all, we obtained 2,708,135 bits of generalized cooccurrence data, which consisted of 115,330 types.

2.3 Measuring the similarity between classes (step 3)

In step 3, we measure the similarity between two primitive classes by using the method given by Hindle (Hindle, 1990).

First, we define the mutual information MI of a verb v and a primitive class C as follows.

$$MI(v,C) = \log_2 \frac{\frac{f(v,C)}{N}}{\frac{f(v)}{N}\frac{f(C)}{N}} \cdots (eq.1)$$

In the above equation, N is the total number of cooccurrence data bits, and $f(v)$ and $f(C)$ are the frequency of v and C in the whole cooccurrence data set respectively, and $f(v,C)$ is the frequency of the cooccurrence data $[C, wo, v]$. Next, the similarity sim of a class C_i and C_j for a verb v is defined as follows.

$$sim(v,C_i,C_j) = \begin{cases} min(|MI(v,C_i)|, |MI(v,C_j)|) \\ \quad : MI(v,C_i) * MI(v,C_j) > 0 \\ 0 \quad : otherwise \end{cases}$$

Finally, the similarity of C_i and C_j is measured as follows.

$$SIM(C_i,C_j) = \sum_v sim(v,C_i,C_j)$$

In equation (eq.1), $f(v) > 0$ because v is the verb in a certain cooccurrence data obtained in step 2. However, $f(C)$ may be equal to 0 because the primitive class C is a certain class in all primitive classes. If $f(C) = 0$, then $MI(v,C)$ cannot be defined. So, if $f(C_i) = 0$ or $f(C_j) = 0$ for all verb v, then $SIM(C_i,C_j)$ is undefined.

2.4 Estimating the undefined similarity (step 4)

There are 3,582 types of primitive classes, so ${}_{3582}C_2 = 6,413,571$ similarities must be defined. Through step 3, there were 2,049,566 similarities which had been defined. This is 32.0 % of the whole.

In step 4, we estimate undefined similarities by the thesaurus and defined similarities. Let us estimate the undefined similarity between the classes C_a and C_b. First, we pick out the set of primitive classes $\{C_{a_1}, C_{a_2}, \cdots, C_{a_i}\}$, such that each class has the common parent node as class C_a in Bunrui-goi-hyou, that is, the class C_{a_i} is the brother node of class C_a. By the same process, we pick out the set of primitive classes $\{C_{b_1}, C_{b_2}, \cdots, C_{b_j}\}$ for class C_b. The similarity in Bunrui-goi-hyou are reliable if its value is large. Thus, it is reliable the defined $SIM(C_{a_k}, C_b)$ and the defined $SIM(C_a, C_{b_m})$ are close to the undefined $SIM(C_a, C_b)$. Therefore, we define $SIM(C_a, C_b)$ by the average of $SIM(C_{a_k}, C_b)$ and $SIM(C_a, C_{b_m})$. This process corresponds to that the slot in the Fig.1(a) is filled with the average of values in the shade part in the figure. If

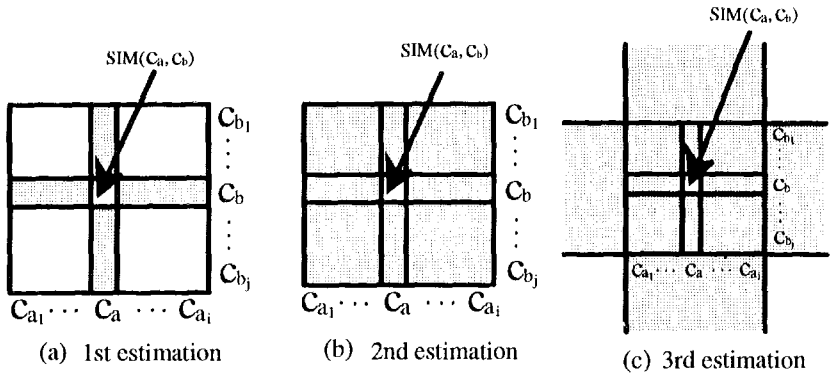

Figure 1: Estimation of $SIM(C_a, C_b)$

the undefined pairs are left through above estimations, they are estimated by the average of $SIM(C_{a_k}, C_{b_m})$. This process corresponds to that the slot in the Fig.1(b) is filled with the average values in the shade part in the figure. If undefined pairs still remain, we pick out the set of primitive classes, such that the grandmother node of each class is the same as that of C_a and C_b , and we repeat the above processes (cf. Fig.1(c)).

Fig.2 shows the ratio of the number of similarities defined in each process.

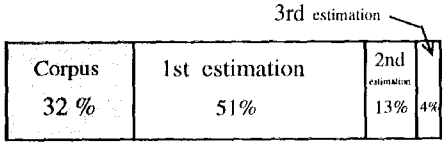

Figure 2: ratio of the number of similarities defined in each process

3 Evaluations

First, we evaluate the obtained similarities by comparing them with the similarities in Bunrui-goi-hyou. The similarity in Bunrui-goi-hyou are defined by the level of the common parent node of two classes. Tab.2 shows the average of similarities between two classes, such that these two classes have the common parent node whose level is x in Bunrui-goi-hyou.

Tab.2 shows that the larger the similarity in Bunrui-goi-hyou is, the larger the obtained similarity is. It follows that the obtained similarity is roughly similar to the similarity in Bunrui-goi-hyou.

Next, we evaluate the appropriateness of the first estimation. The average of "coefficient of variation[1]" for similarities used in each first estimation is 0.384. And the coefficient of variation for all similarities measured by the corpus is 2.125. It follows that similarities used in first estimation are close to each other.

[1]The coefficient of variation is the standard deviation divided by the mean.

the level of the common parent node	average of obtained similarities
1	2.160
2	3.690
3	6.519
4	10.090
5	14.815
6	∞

Table 2: tendency of obtained similarities

At last, we evaluate the appropriateness of the obtained similarity by selecting a verbal meaning. In this experiment, to measure the similarity in Bunrui-goi-hyou and the similarity obtained by our method. Because the similarity in Bunrui-goi-hyou is rough, multiple answers may arise. In evaluation of the similarity in Bunrui-goi-hyou, we give a ○ if the answer is unique and right, a △ if the answers contain the right answer, and × if the answers don't contain the right answer. In evaluation of our similarities, we give a ○ if the largest similarity is right, a △ if 1st or 2nd largest similarities is right answer, and × if neither of 1st and 2nd largest similarities is the right answer.

Tab.1 shows the results of evaluations. This table shows that the similarity obtained by our method is a little better than the similarity in Bunrui-goi-hyou.

4 Remarks

It is difficult to extract all knowledge from only a corpus because of incomplete analysis and data sparseness. In order to avoid these difficulties, the approach to use of different resources from the corpus is promising. To construct the thesaurus from

pattern (num. of meanings)	example nouns	nouns for test	Bunrui-goi-hyou ○	△	×	Our method ○	△	×
を 起こす (9)	27 (妹, 身体, 土, 会社, ...)	24 (看板, 母, 中毒, 水漏れ,...)	14	2	8	17	0	7
が 解ける (4)	12 (ひも, 怒り, 暗号, 処分,...)	4 (結び, しこり, なぞ, 停学)	0	2	2	1	1	2
を 直す (7)	25 (車, ネクタイ, 偏食, 誤字,...)	16 (道, ベンチ, 論文, 結核,...)	8	1	7	9	1	6
を 握る (5)	14 (バット, 包丁, こぶし, 証拠,...)	3 (ひも, マイク, 成否)	1	1	1	1	0	2
に 乗る (5)	18 (電車, 踏み台, 相談, リズム,...)	16 (ロケット, 椅子, 御輿, 人気,...)	14	0	2	13	2	1
に 触れる (3)	16 (彼女, 問題, 核心, 愛,...)	8 (水, 地球, 社会, 制度,...)	7	0	1	7	0	1
を 味わう (3)	13 (幸福, 酒, 古典, 名曲,...)	13 (感触, 余韻, ビール, 作品,...)	12	0	1	10	1	2
を 合わせる (5)	13 (胸, 手, 話, 答え,...)	9 (力, 収入, 曲, つじつま,...)	3	1	5	3	1	5
を 押える (6)	22 (帽子, 目, 怒り, 要点,...)	19 (弦, 犯人, アリバイ, 座席,...)	7	4	8	9	3	7
を 壊す (4)	17 (家, 時計, 腹, 平和,...)	18 (戸棚, 肩, イメージ, 調和,...)	16	0	2	14	2	2
を 練る (2)	6 (計画, 対策, 刀, 飴,...)	8 (策, 構想, 考え, 粉,...)	7	0	1	8	0	0
を 許す (4)	19 (帰宅, 彼, 夜勤, 心,...)	18 (建設, 参加, 利用, 浮気,...)	14	1	3	16	0	2
を 読む (4)	19 (教科書, グラフ, 票, 顔色,...)	28 (解説, 日記, 声明, 動向,...)	13	6	9	16	3	9
Total		184	116	18	50	124	14	46

Table 1: Result of test of verbal meaning selection

a dictionary (Turumaru et al., 1991), and to make example data from a usable knowledge (Kaneda et al., 1995) is considered this approach. The proposed method uses the handmade thesaurus as the different resource from the corpus. In addition, the statistical data from the corpus are weighted. However, it will be important in future research to investigate how much weight should be given to each bit of data.

It is difficult to build knowledge corresponding to each domain from zero. So it is important to extend and modify the existing knowledge corresponding to the purpose of use. In this method, relatively few bits of cooccurrence data are used because nouns in the cooccurrence data are not on Bunrui-goi-hyou. If we extend Bunrui-goi-hyou, these unused cooccurrence data may be useful. And by using the obtained similarities, we can modify Bunrui-goi-hyou. Since our method construct a thesaurus from the handmade thesaurus by the corpus, it can be considered a method to refine the handmade thesaurus such as to be suitable for the domain of the used corpus.

5 Conclusions

In this paper, we proposed a method to define similarities between general nouns used in various domains. The proposed method redefines the similarity in a handmade thesaurus by using corpora. The method avoids data sparseness by estimating undefined similarities from the similarity in the thesaurus and similarities defined by corpora. The obtained similarities are obviously the same in number as the original similarities, and are more appropriate than the original similarities in the thesaurus.

By using Bunrui-goi-hyou as the handmade thesaurus and newspaper articles with about 7.85 M sentences as a corpus, we confirmed the appropriateness of this method.

In the future, we will extend and modify Bunrui-goi-hyou by the cooccurrence data and the similarities obtained in this study, and will try to classify multiple senses of verbs.

Acknowledgment

The corpus used in our experiment is extracted from CD-ROMs ('90 - '94) sold by Nihon Keizai Shinbun company. We deeply appreciate the Nihon Keizai Shinbun company to permit the use of this corpus and many people who negotiated with the company about the use of this corpus.

References

Brown,P.F.,Pietra,V.D, deSouza,P.V., Lai,J.C. and Mercer,R.L. : 1992. Class-Based n-gram Models of Natural Language, Computational Linguistics, Vol.18,No.4,pp.467–479(1992).

Dagan,I., Marcus,S., and Markovitch,S. : 1993. Contextual Word Similarity and Estimation from Sparse Data, In *31th Annual Meeting of the Association for Computational Linguistics*, pp.164–171.

Dagan,I., Pereira,F., and Lee,L. : 1994. Similarity-Based Estimation of Word Coocurrence Probabilities, In *32th Annual Meeting of the Association for Computational Linguistics*, pp.272–278.

Hindle,D. 1990. Noun classification from predicate-argument structures. In *28th Annual Meeting of the Association for Computational Linguistics*, pp.268–275.

Kaneda,S., Akiba,Y., and Ishii,M. : 1995. Jireini motozuku eigodousi sentakuruuru no syuuseigata gakusyuuhou (in Japanese), In *Proceedings of the first annual meeting of the Assoiation for Natural Language Processing*, pp.333–336.

Pereira,F., Tishby,N., and Lee,L. : 1993. Distributional Clustering of English Word, In *31th Annual Meeting of the Association for Computational Linguistics*,pp.183–190.

The National Language Research Institute : 1994. Bunrui-goi-hyou (in Japanese), Shuuei Publishing.

Turumaru,H., Takesita,K., Itami,K., Yanagawa,T. and Yoshida,S. : 1991. An Approach to 'Thesaurus Construction from Japanese Language Dictionary (in Japanese), IPS Japan NL-83-16, Vol.91,No.37,91-NL-83, pp.121–128.