情 報 学 基 礎 61－9
自 然 言 語 処 理 142－9
（2001. 3. 5）

Determination of the Meaning of Polysemous Words Using a Word Similarity Measurement

Qujiang Peng Sawa Takakura Teiji Furugori
Department of Computer Science
University of Electro-Communications
1-5-1, Chofugaoka, Chofu, Tokyo 1828585, JAPAN
{peng, furugori}@phaeton.cs.uec.ac.jp

Abstract

We describe a method and its experimental results for word sense disambiguation that is based on a statistical measure of word similaritites. First, we obtain contextual-similarity vectors for the senses of a polysemous word using a corpus. Second, we define also the contextual representation for the same word appearing in text. Third, we do a calculation of distributional matrix between each contextual-similarity vector and the contextual representation for the word to be disambiguated. Fourth and finally, comparing the values of distributional matrices, we select the sense with the highest value as the meaning of the polysemous word. An experiment shows that the rate of finding correct word senses exceeds over 91%.

Keywords: word sense disambiguation, polysemous words, contextual similarity, distributional matrix, density

単語の類似性の尺度を使った多義語の意味の決定法

彭 渠江 高倉 佐和 古郡 廷治
〒182-8585 東京都調布市調布ケ丘 1－5－1, 電気通信大学情報工学科

概要

本稿では、単語の意味的曖昧性を解く手法の開発と、それをもとにして行った曖昧性解消の実験結果を報告する。テキスト中の単語の語義（sense）は、一定の文脈の中で、その単語とよく共起する他の単語と高い相互情報量をもつ。この特徴を使い、単語（*w*）が使われている文脈中で出現し、*w*と類似度の高い単語のベクトルと、*w*がもつ *r* 個の語義のそれぞれが使われている文脈中で出現し、*w*と類似度の高い *k* 個の単語のベクトルとの間の相互情報量を計算し、その値が最も高くなった密度値と結合している語義を*w*の語義として採用する。この手法によって行った実験では、９１．５％の高率で多義語の正しい語義を特定することができた。

キーワード: 語義的曖昧性の解消、多義語、文脈的類似性、分布的マトリックス、密度

1 Introduction

Ambiguities observed on all levels of language are the problem in natural language processing. Typically they are noticed in word meanings. For instance, it is impossible to translate the following text into Japanese unless we get exact meaning of *sentence* that has at least two meanings: a group of words or punishment.

Taro got a heavy sentence for the crime he committed.

In this paper, we present a method and its experimental results for resolving lexical ambiguities. We base our method on a word similarity measure that uses the mutual information.

2 Word Similarity

We follow in the footsteps of similarity-based approach to form our method for word sense disambiguation (WSD). We start with obtaining contextual-similarity vector for each sense of a polysemous word and the contextual representation for the same word appearing in text. We then calculate distributional matrix between each vector and the contextual representation. Finally, we compare the values of distributional matrices and select the sense with the highest value as the meaning of the polysemous word in question.

Mutual Information and Similarity Metric of Two Words We use mutual information to get the contextual-similarity vector and the distributional matrix for each sense of a polysemous word. The mutual information, *I*, estimates the strength of association between two words w_1 and w_2 [2]:

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

Here, *N* in (1) is the size of the corpus used in the estimation, $f(w_1, w_2)$ is the frequency of co-occurrences of w_1 and w_2, and $f(w_1)$ and $f(w_2)$ is the frequency of each word. If there is a strong association between w_1 and w_2, then $I(w_1, w_2) >> 0$. If there is a weak association between w_1 and w_2, then $I(w_1, w_2) \approx 0$. If $I(w_1, w_2) << 0$, then w_1 and w_2 are said to be in complementary distribution.

The contextually similar words are the words co-occurring frequently in a distance in text. They do not need to be synonyms, or belong to the same syntactic or semantic category. For instance, the words *doctor* and *health* are far removed in a typical semantic hierarchy being defined (e.g., WordNet [8]), but they are said to be contextually similar as they tend to co-occur in a text.

Dagan, et al. [3] use mutual information to define the contextual-similarity of two words as:

$$sim(w_1, w_2) = \frac{\sum_{w \in Lexicon} \min(I(w, w_1), I(w, w_2)) + \min(I(w_1, w), I(w_2, w))}{\sum_{w \in Lexicon} \max(I(w, w_1), I(w, w_2)) + \max(I(w_1, w), I(w_2, w))} \qquad (2)$$

An assumption here is that w_1 and w_2 have similar mutual information with some other word w if the two words are similar.

By observation, Dagan, et al. find, "when computing $sim(w_1, w_2)$, words with high mutual information values with both w_1 and w_2 make the largest contributions to the value of the similarity measure. Also, high and reliable mutual information values are typically associated with relatively high frequencies of the involved co-occurrence pairs. "

A topical sense of a polysemous word occurs in a topical context. That is to say, the topical sense of a polysemous word co-occurs with certain words in the topical context: the polysemous word used in the topical sense has high mutual information values with the certain words. When computing the similarity of a topical sense of a polysemous word, *s*, and a word, *y*, in the same topical context, some words in the context make the largest contributions to the value of the similarity measure because only they have chance to have high mutual information values with both *s* and *y*.

If a word w_2 usually co-occurs with some words, and these words co-occur with a topical sense of a polysemous word, *s*, at the same time, then w_2 is likely to become similar word of *s*. Especially, a word is most similar to itself.

Dagan, et al. give a heuristic algorithm to search for the *k* most similar words. We use their method to find contextually similar words of each sense of a polysemous word, *s*. Each sense of the polysemous word and its similar word have similar mutual information with some other words. If a word w_2 does not co-occur with words in the context of a topical sense of a polysemous word, *s*, then w_2 is unlikely to become a similar word of *s*, because *s* co-occurs almost only with words in its topical context. So similar word of *s* has co-occurrence with certain words in the topical context. A topical sense of a polysemous word, s, and its similar word have similar mutual information with some other words in its topical context. We use this feature in our disambiguation method.

Contextual Representation and Contextual-Similarity Vector Word appears in a context. The contextual representation (CR) we use here is the sequence of content words (nouns, verbs, adjectives) that appear in a distance l with the polysemous word in question. It is defined as the vector:

$$V_{CR} = (w_1, w_2, \cdots, w_n) \qquad (3)$$

Miller and Charles [9] found evidence in several experiments that humans determine the semantic similarity of words from the similarity of the contexts the words are used in. Extending the finding, Schütze [10] hypothesized that the same holds for word senses: a sense is a group of contextually similar occurrences of a word. Karov and Edelman [5] used that words are considered similar if they appear in similar contexts and contexts are similar if they contain similar words.

Let s_m be the m th sense of a polysemous word w. By definition (2) and the heuristic algorithm Dagan, et al. used, we can get a vector of each sense of a polysemous word.

$$V_{s_m} = \left(w_1^m, w_2^m, \cdots, w_k^m\right) \qquad (4)$$

Here $w_1^m, w_2^m, \cdots, w_k^m$ are the set of the k most contextually similar words of the m th sense of a polysemous word w. s_m and w_i^m $(i = 1, \cdots, k)$ have similar mutual information with some other word. We use $k = 6$ in this paper.

When the m th sense of a polysemous word w in the corpus faces the data sparseness problem, we seek its synonym set in the WordNet [13], choose from the set a word that does not have the problem, and use it to calculate V_{s_m}. When no such word is found, we use its coordinate word that does not have the problem to calculate V_{s_m}.

3 Distributional Matrix and the Meaning of Polysemous Word

Using (3) and (4), we can get a distributional matrix of mutual information

$$M\left(V_{s_m}, V_{CR}\right) = \left(I\left(w_i^m, w_j\right)\right)_{k\times n} \qquad (5)$$

Here $i = 1, \cdots, k$ and $j = 1, \cdots, n$. The matrix $M\left(V_{s_m}, V_{CR}\right)$ expresses a distribution of the tight degree of association between the m th sense of a polysemous word w and the context V_{CR}. And we can get the density of the matrix $M\left(V_{s_m}, V_{CR}\right)$.

$$\rho\left(V_{s_m}, V_{CR}\right) = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n} I\left(w_i^m, w_j\right)}{k \times n} \qquad (6)$$

The idea in (6) comes from the definition of density in Physics. The density in Physics expresses a distribution of substance in one volume. Essentially, it expresses a tight degree of association among the corpuscles of substance. If there is a strong association among the corpuscles of substance, then its density is big. For example, the density of solid is bigger than the density of gas because the movement of the corpuscles of solid is not free than that of gas. The corpuscle of solid is fastened by the other corpuscles around. It means that the tight degree of association among the corpuscles of solid is very strong. Here, w_i^m and w_j are equivalent to the corpuscles of substance, and $I\left(w_i^m, w_j\right)$ is equivalent to the tight degree of association between the corpuscles of substance.

In the context of natural language, it is natural to say that there are interferential words or irrelevant words in V_{CR} to express the topic involved. So we set the threshold t to modify (6). In the matrix $M(V_{s_m}, V_{CR})$, we throw out the values of mutual information in d most small values if the value is less than t. For the rest, we calculate the density of mutual information. We mark the density with $sense(V_{s_m}, V_{CR})$. The parameters d and t are adjustable.

In our experiment, we use windows of l=10, 35, and 50 words before and after the polysemous word to be disambiguated. When $l=10$, we set none of t and d. For $l=35$ and $l=50$, we set t to be

$$t = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n} I(w_i^m, w_j)}{4k \times n}$$

and d to be *9k* and *12k*, respectively.

Suppose the word appearing in a text is w, and its lexical meanings are $s_1, s_2, \cdots, s_r$. We calculate the density $sense(V_{s_1}, V_{CR})$, $sense(V_{s_2}, V_{CR})$, $\cdots$, $sense(V_{s_r}, V_{CR})$. The meaning of the word in consideration is s_m that satisfies the maximal value.

4 Experiment and Results

To test our method, we use EDR English Corpus [12] as training data and 10 polysemous words for which 682 instances were selected randomly from [14, 15] and other materials. The EDR corpus contains 160,000 sentences with annotated morphological, syntactic and semantic information.

We predetermined the meaning of each instance of the polysemous word by two human subjects. No directionality in (2) is assumed, i.e., $(w, w_1) = (w_1, w)$ and $(w, w_2) = (w_2, w)$. The meaning of w is determined by the following procedure:

(1) Obtain the vectors $V_{s_1}, V_{s_2}, \cdots, V_{s_r}$ for the lexical meanings of w.

(2) Get V_{CR} in consideration using a window of l words each before and after w in the text.

(3) Calculate the density $sense(V_{s_1}, V_{CR})$, $sense(V_{s_2}, V_{CR})$, $\cdots$, $sense(V_{s_r}, V_{CR})$.

(4) Select s_m that got the maximal density value to be the meaning of w.

Examples Consider the polysemous word in consideration to be *cabinet* in the following text and see how its meaning is determined:

> ··· Chicago was the industrial inferno of the nineteenth century A.D. A curious anecdote has come down to us of John Burns, a great English labor leader and one time member of the British *cabinet*. In Chicago, while on a visit to the United States, he was asked by a newspaper reporter for his opinion of that city. 'Chicago,' he answered, 'is a pocket edition of hell.' Some time later ··· (*a text on Internet*)

Cabinet is given two nominal meanings:

administrative organ (s_1) and a shelf (s_2). Using the method by Dagan, et al., we get for them the contextual-similarity vectors:

V_{s_1} = *(approve, minister, formal, meeting, conference, submit)*

V_{s_2} = *(exhibit, maker, store, clothes, lock, shelf)*

The contextual representation V_{CR} ($l = 35$) from the text is:

V_{CR} = *(⋯, time, member, British, Chicago, visit, ⋯)*

The calculation of the distributional matrix produces the density 0.346237 for $sense(V_{s_1}, V_{CR})$ and 0.124824 for $sense(V_{s_2}, V_{CR})$. So, we decide the meaning of the *cabinet* to be s_1 (administrative organ) .

Take another example for the word *sentence* ($l = 35$) in:

> ⋯ and there he died like a brave man. He refused to have his eyes bandaged, saying that he was not at all afraid of death; and he admitted the justice of his *sentence*, and was much regretted by the people. Although Mary had shrunk at the most important time from disproving her guilt, she was very careful never to do anything that would admit ⋯ (*Also a text on Internet*)

Again, *sentence* has two meanings: punishment (s_1) and group of words (s_2). The contextual-similarity vectors and the contextual representation are:

V_{s_1} = *(prison, convict, defendant, jail, imprisonment, guilty)*

V_{s_2} = *(meaning, booklet, language, word, dictionary, accent)*

V_{CR} = *(⋯, death, admit, justice, regret, people, ⋯)*

Now, the density for $sense(V_{s_1}, V_{CR})$ becomes 0.353611 and $sense(V_{s_2}, V_{CR})$ becomes 0.130465. Thus, we get the meaning of the *sentence* as s_1 (punishment).

Results Table 1 contains the 10 polysemous words, their senses, and the number of their instances. Table 2 shows the disambiguation result (success rate). The results with l=10, 35, and 50 are shown for Experiment 1, Experiment 2, and Experiment 3, respectively.

Table 1: Polysemous Words Tested

Words	Senses	Instances
band	group of musicians strips or stripes	19 4
cabinet	administrative organ shelf	24 17
court	judicial area for ball game	163 9
crane	machine bird	16 21
palm	tree hand	11 52
plant	living thing factory	86 18
sentence	group of words punishment	25 67
slug	bullet animal	6 16
tank	combat vehicle water-filled place	12 10
trial	action of judging test	89 17

Table 2: Disambiguation Result(%)

Words Test	band	cabinet	court	crane	palm	plant	sentence	slug	tank	trial	Average
Experiment 1	87.0	78.0	90.7	91.9	90.5	92.3	88.0	77.3	95.5	89.6	89.3
Experiment 2	95.7	87.8	95.9	91.9	85.7	90.4	82.6	90.9	90.9	88.7	90.2
Experiment 3	95.7	97.6	95.3	97.3	85.7	91.3	81.5	72.7	90.9	96.2	91.5

Evaluation As is seen in Table 2, the best result is got for *cabinet* when $l = 50$ and the worst is for *slug* when $l = 50$. The overall average success rates are 89.3%, 90.2%, and 91.5%, respectively, for Experiment 1, Experiment 2, and Experiment 3. There are some cases where we got one side of meaning all right and the other side considerably worse, e.g., *palm* and *tank*. We see for some words that the success rates vary very much depending on the window sizes.

Comparative evaluation is generally difficult for word sense disambiguations due to the differences in detailed methodologies and test data used. Percentage-wise, however, our success rate is mostly better than the ones in other studies [1, 4, 11], and much better than the performance (72% in disambiguating nouns) of the WordNet-based method proposed by Li, et al. [7]

5 Conclusion

We proposed a method for WSD that uses the contextual-similarity vector and the idea of density in Physics for resolving ambiguities of polysemous words in text.

Our method is intuitional and its performance is in an acceptable level. However, to improve the performance, it may be worthwhile to consider such points as:

- Dynamic adjustment of parameters d and t for extracting the topic effectively from the context.
- Invertion of the distributional matrix to distributional image in the two dimensional coordinates, and the determination of the sense according to distributional character of the image.

In our method and many others, context is the means to identify the meaning of a polysemous word. But some word senses are nonspecific to topics and a word with such a sense can appear freely in many domains of discourse. It is desirable to give a thought to deal with the meaning of polysemous words used in the nontopical sense [6]. However, within a restriction, we claim that our method is theoretically sound and better in generality and flexibility.

References

[1] Chen, J. N. & Chang, J. S., A Concept-based Adaptive Approach to Word Sense Disambiguation, In *Proceedings of*

COLING-ACL '98, 237-243, Montreal, Quebec, Canada (1998)

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

[3] Dagan, I., Marcus, S & Markovitch, S., Contextual Word Similarity and Estimation from Sparse Data, *Computer Speech and Language*, 9, 123-152 (1995)

[4] Hiro, K., Wu, H. & Furugori, T., Word-Sense Disambiguation with a Corpus-Based Semantic Network, *Journal of Quantitative Linguistics*, 3, 244-251 (1996)

[5] Karov, Y. & Edelman, S., Similarity-based Word Sense Disambiguation, *Computational Linguistics*, 24(1), 41-59 (1998)

[6] Leacock, C., Chodorow, M. & Miller, G. A., Using Corpus Statistics and WordNet Relations for Sense Identification, *Computational Linguistics*, 24(1), 147-165 (1998)

[7] Li, X., Szpakowicz, S. & Matwin, S., A WordNet-based Algorithm for Word Sense Disambiguation, In *Proceedings of IJCAI-95*, 1368-1374, Montreal: Morgan Kaufmann (1995)

[8] Miller, G. A., Richard, B., Christiane, F., Derek, G. & Katherine, J. M., *Introduction to WordNet: An On-line Lexical Database*, CSL 43, Cognitive Science Laboratory, Princeton University, Princeton, NJ (1993)

[9] Miller, G. A. & Walter G. Charles, Contextual Correlates of Semantic Similarity, *Language and Cognitive Processes*, 6(1), 1-28 (1991)

[10] Schütze, H., Automatic Word Sense Discrimination, *Computational Linguistics*, 24(1), 97-123 (1998)

[11] Yarowsky, D., Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora, In *Proceedings of COLING-92*, 454-460, Nantes:ICCL (1992)

[12] http://www.iijnet.or.jp/edr/index.html

[13] http://www.cogsci.princeton.edu/~wn/

[14] http://gamp.c.u-tokyo.ac.jp/archive/textdb.htm

[15] http://www.infomotions.com/etexts/