Automatic Word Sense Clustering Using Collocation for Sense Adaptation

Sa-Im Shin and Key-Sun Choi

KORTERM, KAIST
373-1 Guseong-dong, Yuseong-gu, Daejeon, Republic of Korea
Email: miror@world.kaist.ac.kr, kschoi@world.kaist.ac.kr

Abstract. A specific sense of a word can be determined by collocation of the words gathered from the large corpus that includes context patterns. However, homonym collocation often causes semantic ambiguity. Therefore, the results extracted from corpus should be classified according to every meaning of a word in order to ensure correct collocation. In this paper, K-means clustering is used to solve this problem. This paper reports collocation conditions as well as normalized algorithms actually adopted to address this problem. As a result of applying the proposed method to selected homonyms, the optimal number of semantic clusters showed similarity to those in the dictionary. This approach can disambiguate the sense of homonyms optimally using extracted texts, thus resolving the ambiguity of homonyms arising from collocation.

1 Introduction

In a wide sense, collocation is a pattern of words that coexist in the fixed window or in the same sentence. According to the Yahoo monolingual dictionary, a Korean word *shinbu* has five senses: (1) believable or unbelievable work; (2) a certificate in old Korea – Chosun; (3) a Catholic priest; (4) an amulet; (5) a bride. So, the collocation with the word *shinbu* results in ambiguous context information due to the five senses.?? Through the collocations extracted from the corpus, we can seize some facts concerning the collocation. For example, 'Buddhist priest', 'Africa', 'Braman', 'discipline', and 'appointment', etc. are related to the third sense 'Catholic priest', while 'couple', 'beautiful', 'match', etc. are collocations for the fifth sense 'bride'.

This paper is organized as follows. Firstly, we introduce the sense clustering model of collocation and discuss how to decide the optimal number of cluster. Secondly, we suggest the similarity measure in terms of validity. Finally, we observe and discuss on the experimental results comparing them with other researches.

2 Representation of Collocation

The words in the collocation also have their collocations. A target word for collocation is called the 'central word', and a word in a collocation is referred to as the 'contextual word'. Upon the assumption that there are $w$words placed in the right and left of the central word x, contextual words $x_i^{\pm j}$ for the central word x are represented as follows:

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 320–325.
© Masaryk University, Brno, 2003

$\langle x_i^{-w}, \ldots, x_i^{-1}, x, x_i^{+1}, \ldots, x_i^{+w} \rangle$. If collocation patterns between contextual words are similar, it means that the contextual words are used in a similar context – where used and interrelated in same sense of the central word – in the sentence. If contextual words are clustered according to the similarity in collocations, contextual words for polysemous central words can be classified according to the senses of the central words.

The following is a mathematical representation used in this paper. A collocation of the central word x, window size w and corpus c is expressed with function $f : V \times N \times C \rightarrow 2^{C/V}$. In this formula, V means a set of vocabulary, N is the size of the contextual window that is an integer, and C means a set of corpus. In this paper, vocabulary refers to all content words in the corpus. Function f shows all collocations. C/V means that C is limited to V as well as that all vocabularies are selected from a given corpus and $2^{C/V}$ is all sets of C/V. In the equation (1), the frequency of x is m in c. We can also express $m = |c/x|$. The window size of a collocation is $2w + 1$.

$$f(x, w, c) = \left\{ \begin{array}{c} \langle x_1^{-w}, \ldots, x_1^{-1}, x, x_1^{+1}, \ldots, x_1^{+w} \rangle \\ \ldots \\ \langle x_m^{-w}, \ldots, x_m^{-1}, x, x_m^{+1}, \ldots, x_m^{+w} \rangle \end{array} \right\} \tag{1}$$

$g(x) = \{(x, i), i \in I_x\}$ is a word sense assignment function that gives the word senses numbered i of the word x. I_x is the word sense indexing function of x that gives an index to each sense of the word x. All contextual words $x_i^{\pm j}$ of a central word x have their own contextual words in their collocation, and they also have multiple senses. This problem is expressed by the combination of g and f as follows:

$$g \circ f(x, w, c) = \left\{ \begin{array}{c} \langle g(x_1^{-w}), \ldots, g(x_1^{-1}), g(x), g(x_1^{+1}), \ldots, g(x_1^{+w}) \rangle \\ \ldots \\ \langle g(x_m^{-w}), \ldots, g(x_m^{-1}), g(x), g(x_m^{+1}), \ldots, g(x_m^{+w}) \rangle \end{array} \right\} \tag{2}$$

In this paper, the problem is that the collocation of the central word is ordered according to word senses.

3 Sense Clustering Model Using Collocation

This research applies K-means clustering [4] to the automatic clustering of collocations as introduced below. This method classifies contextual words of the central word into K clusters. For this method, |Ix| refers to K. This approach has been used to extract collocations within a similar context and sense of the central word.

1. Choose K initial cluster centers $z_1(1), z_2(1), \ldots, z_K(1)$, where $k = 1$.
2. At the k-th iterative step, distribute the corpus $\{x\}$ among K clusters by the following condition, where $C_j(k)$ denotes a cluster whose center is $z_j(k)$:

$$x \in C_j(k) \text{ if } sim(x, z_j(k)) > sim(x, z_i(k)), \quad i = 1, 2, \ldots, K; \quad i \neq j \tag{3}$$

1. Compute a series of new cluster centers $z_j(k + 1)$, $j = 1, 2, \ldots, K$ in a way that minimize the sum of similarities from all points in $C_j(k)$ to the new cluster center.

2. If $||z_j(k+1)-z_j(k)||<\alpha$ ($j = 1, 2, \ldots, K$), then terminate. Otherwise, go to step 2.

Corpus c is represented as $\{x_i\}$ $(1\leq i\leq q)$:q is the number of unique words in c). $(t_{i1}, \ldots, t_{iq})$ is a vector representation of each word x_i according to contextual words as well as co-occurrence frequency. $(t_{i1}, \ldots, t_{iq})/w$ is represented on account of restrictions by the fixed window size w. A cluster C_{aj} means the j-th cluster of the central word x_a. The center z_{aj} of each cluster C_{aj} and the contextual word x_a^i $(i = -w, \ldots, +w)$ for x_a is represented as follows:

$$\vec{z}_{aj} = \left(t_1^{z_{aj}}, \ldots, t_q^{z_{aj}}\right), \quad \vec{x}_a^i = \left(t_1^{x_a^i}, \ldots, t_q^{x_a^i}\right) \tag{4}$$

Each frequency is $t_a^b = \log(P_{ab}/P_a P_b)$ while a and b are targets for collocation. P_{ab} is the probability of co-occurrence between a and b. The cosine similarity between the center z_{aj} and the contextual word x_a^i is expressed as follows:

$$sim\left(\vec{z}_{aj}, \vec{x}_a^i\right) = \sum\nolimits_{m=1,q} t_m^{z_{aj}} t_m^{x_a^j} \Big/ \sqrt{\sum\nolimits_m (t_m^{z_{aj}})^2 \sum\nolimits_m (t_m^{x_a^j})^2} \tag{5}$$

During this process, each contextual word is classified into one cluster having the largest similarity value while repeating this process until the results remain unchanged.

4 Sense Clustering Algorithm: Similarity and Optimal Decision

During the k-th repetition, we update a new center $z_j(k+1)$ using an average of the newly generated j-th cluster $C_j(k)$ based on the center $z_j(k)$. Revised K-means clustering algorithm for sense clustering is addressed by the following subsections.

4.1 Determination of Cluster Centers

In the process of initial clustering, the centers of clusters are determined by randomly selected K contextual words. In each clustering cycle, their centers are adjusted by the average frequency of the contextual words in each cluster. Throughout the repetition process, the center of each cluster is converged to the real center, and similar contextual words are also clustered toward these centers. The equation (6) shows the center $z_j(k+1)$ of the j-th cluster for the next clustering cycle.

$$\vec{z}_j(k+1) = \left(\tfrac{1}{N_j}\textstyle\sum_{i=1,q} t_i^1, \ldots, \tfrac{1}{N_j}\sum_{i=1,q} t_i^q\right) \tag{6}$$

4.2 Termination Conditions for Clustering

The clustering algorithm cycle repeats until the clustering results become stable. It determines whether termination requirements are met. If clustering results meet termination requirements without any change in the clustering results, the clustering process is completed. In this paper,

termination conditions are determined by the rate of variations after each clustering cycle. The following equations indicate the validity in the p-th clustering cycle.

$$\begin{aligned} validity_p &= intra_p \big/ inter_p \\ intra_p &= 1/N \sum_{i=1,K} \sum_{x \in C_i} sim(\vec{x}, \vec{z}_i) \\ inter_p &= \max(sim(\vec{z}_i, \vec{z}_j)) \end{aligned} \tag{7}$$

Internal cohesion $intra_p$ is the average similarity between the center of each cluster and its members, which measures the cohesion of each cluster. External similarity $inter_p$ is the maximum similarity among the centers of clusters and this value is determined by the similarity with the most similar cluster. So, external similarity expresses the discrimination of the clusters. If variations of the validity are lower than that of the threshold, the clustering process is completed. Our experiments show the threshold is 10^{-6}.

5 Experiment and Analysis

5.1 Collocation Normalization

In order to apply correct collocation, it need to remove the noise and trivial collocation. This process is called normalization, and its process is specifically provided as follows: (1) remove noise in the tagging or the corpus; (2) remove the words of foreign origin – aimed at avoiding data sparseness; (3) remove one-syllable words; (4) remove statistically unrelated words. According to the Zipf's law, 80% of the words appear only once, while the rest forms 80% of the corpus [1]. Therefore, it can be said that the words with high frequency appear regardless of their semantic features [1]. High frequency words like this are called statistically unrelated words. The words with high frequency can be removed not only by these statistically unrelated words through the sorting of collocations of each contextual word but also by frequently appearing words.

5.2 Number of Clusters

In this research, the number of cluster K is not arbitrarily determined. K refers to the ambiguity of the central word. Therefore, it is important to determine K in reflecting the real ambiguity of the central word. In the process of performing repeated experiments, we selected the optimal number of clusters according to the ambiguity of the central word. The variance of K is determined by statistical analysis of existing dictionaries. The result of analysis of nouns, adjectives, verbs and adverbs in [8] shows that the word with the maximum number of meanings has 41 senses, except for frequently appearing one-syllable words that we already removed.

5.3 Experimental Results

We used KAIST corpus for the experiments [7]. We extracted collocations of about 10 million words from the KAIST corpus. In the experiments, K-means clustering was applied to some of the most famous homonyms. The clustering results are shown in Table 1. The "normalizing rate" means the rate of removing statistically unrelated words in a collocation

Table 1. Clustering Results by Each Normalizing Rate

Normalizing rate / Number of clusters			*A*	*B*	*C*	*D*
word	[8]	[9]				
shinbu	5	2	2	14	9	10
yuhag	5	3	2	2	6	6
buja	4	3	2	3	5	6
sudo	2	2	2	2	4	4
gong'gi	5	5	2	4	4	5

of each contextual word. A indicates the results without normalization, while B, C, and D indicate the results of clustering at a normalizing rate of 0%, 30%, and 50%, respectively.

The results of clustering show that the words are classified in a more specific way than in dictionaries. However, for unnormalized clusterings as in A, the number of clusters is smaller than that shown in other results. That's because correct clustering was interfered by the noise as well as most of frequently appearing words in the collocation. But B, C and D sometimes construct unsuitable clusters. If an initial center is determined by most frequently appearing words, many contextual words are clustered in this cluster. Because most frequently appearing words contain richer context, this center is most likely to match than less frequently appearing centers.

Fluctuations in the number of clusters are found similar to the word senses in dictionaries. It means that K-means clustering proposed in this paper ensures achieving the optimal number of clusters K. Research results show the distribution of practical senses appearing in the corpus.

6 Results and Discussion

This paper is intended to resolve sense ambiguity in collocations through the removal of the noise as well as the application of an automatic clustering method – K-means clustering. Since the clustering method proposed in this paper determines automatically the number of clusters based on the corpus, these numbers guarantee optimal clustering results that have led us to extract practical senses in conducting our research.

However, this research suggests some points to further improve. Firstly, the collocation of contextual words that is used to disambiguate also contains ambiguity. This recursive ambiguity is still reflected on the results. In the second place, the pattern of contextual words in Equation (1) evolves into sorting and clustering problems when contextual words are expressed with their specific senses like Equation (2). This problem is also clearly dependent on how to represent each sense set $g(x)$; the one definitions in dictionaries or logical forms, the other by semantic categories in thesaurus. We need to integrate this research into the research on thesaurus and sense definition. In the third place, there is a problem concerning the optimality of clustering. The method of calculating similarity is affected by clustering results. In this paper, similarity is determined by the relative frequency of occurrence of the collocation of contextual words. Since calculating similarity is accompanied by a judgment which contextual words can be effective in performing clustering, the ensuing experiments

must be performed. Specifically, the one is the clustering method using LSI (Latent Semantic Indexing), based on the frequency of occurrence of words, and the other approach to clustering is using the classical *tf-idf* method. This paper applies the latter method indirectly, but more direct application is needed.

Heyer, et al. [5,6] extracted collocations from a large-scaled corpus and constructed an online dictionary called 'Wortschatz'. Nonetheless, this work contains normalizing problems concerning the occurrence pattern of collocations. Lin [3] constructed an English thesaurus using an automatic clustering. But the thesaurus is not only locally and limitedly covered, but also are sense ambiguities in the words excluded. Park [2] constructed a collocation map using collocation and Bayesian network. Pantel [10] discovered senses from English text using CBC and proposed the evaluation measure of this result by comparing with WordNet [12]. This method considered limited contextual words in [10] – frequent nouns over the threshold, but we allow all content words in the same sentence. Ji [11] proposed the *clinique* and clustering methods for collocation clustering. [11] applied collocation clustering in the sense ambiguity in machine translation – selecting translated words and sense ambiguity in compound nouns.

References

1. Zipf G.: Human Behavior and the Principle of Last Effort. Cambridge (1949).
2. Young C. Park and Key-Sun Choi: Automatic Thesaurus Construction Using Bayesian Networks. Information Processing and Management (1996).
3. Lin D.: Automatic Retrieval and Clustering of Similar Words. Coling-ACL (1998).
4. Ray S., Turi R. H.: Determination of Number of Clusters in K-means Clustering and Application in Colour Image Segmentation. In Proceedings of the 4th International Conference on Advances in Pattern Recognition and Digital Techniques, India (1999).
5. Heyer G., Quasthoff U., Wolff C.: Information Extraction from Text Corpora. IEEE Intelligent Systems and Their Applications (2001).
6. Läuter M., Quasthoff U., Wittig T., Wolff C., Heyer G.: Learning Relations using Collocations. IJCAI (2001).
7. KAIST Corpus: `http://kibs.kaist.ac.kr/` (1999–2003).
8. Hangeul Society, ed.: Urimal Korean Unabridged Dictionary, Eomungag, (1997).
9. Yonsei Dictionary: `http://clid.yonsei.ac.kr:8000/dic/default.htm` (2003).
10. Patrick Pantel and Dekang Lin.: Discovering Word Senses from Text. In Proceedings of ACM Conference on Knowledge Discovery and Data Mining. Edmonton, Canada. (2002).
11. Hyungsuk Ji, Sabine Ploux and Eric Wehrli.: Lexical Knowledge Representation with Contexonyms. In Proceedings of the 9th Machine Translation. (2003).
12. WordNet: `http://www.cogsci.princeton.edu/~wn/`.