Automatic clustering of collocation for detecting practical sense boundary

Saim Shin
KAIST
KorTerm
BOLA
miror@world.kaist.ac.kr

Key-Sun Choi
KAIST
KorTerm
BOLA
kschoi@world.kaist.ac.kr

Abstract

This paper talks about the deciding practical sense boundary of homonymous words. The important problem in dictionaries or thesauri is the confusion of the sense boundary by each resource. This also becomes a bottleneck in the practical language processing systems. This paper proposes the method about discovering sense boundary using the collocation from the large corpora and the clustering methods. In the experiments, the proposed methods show the similar results with the sense boundary from a corpus-based dictionary and sense-tagged corpus.

1 Introduction

There are three types of sense boundary confusion for the homonyms in the existing dictionaries. One is sense boundaries' overlapping: two senses are overlapped from some semantic features. Second, some senses in the dictionary are null (or non-existing) in the used corpora. Conversely, we have to generate more senses depending on the corpora, and we define these senses with practical senses. Our goal in this study is to revise sense boundary in the existing dictionaries with practical senses from the large-scaled corpus.

The collocation from the large-scaled corpus contains semantic information. The collocation for ambiguous words also contains semantic information about multiple senses for this ambiguous word. This paper uses the ambiguity of collocation for the homonyms. With the clustering algorithms, we extract practical sense boundary from the collocations.

This paper explains the collocation ambiguity in chapter 2, defines the extracted collocation and proposes the used clustering methods and the labeling algorithms in chapter 3. After explaining the experimental results in chapter 4, this paper comes to the conclusion in chapter 5.

2 Collocation and Senses

2.1 Impractical senses in dictionary

In (Patrick and Lin, 2002), senses in dictionary – especially in WordNet – sometimes don't contain the senses appearing in the corpus. Some senses in the manual dictionary don't appear in the corpus.

This situation means that there exist differences between the senses in the manual dictionaries and practical senses from corpus. These differences make problems in developing word sense disambiguation systems and applying semantic information to language processing applications.

The senses in the corpus are continuously changed. In order to reflect these changes, we must analyze corpus continuously. This paper discusses about the analyzing method in order to detect practical senses using the collocation.

2.2 Homonymous collocation

The words in the collocation also have their collocation. A target word for collocation is called the 'central word', and a word in a collocation is referred to as the 'contextual word'. 'Surrounding words' mean the collocation for all contextual words. The assumption for extracting sense boundary is like this: the contextual words used in the same sense of the central word show the similar pattern of context. 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 homonymous central words can be classified according to the senses of the central words. (Shin and Choi, 2004)

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: $V\ N\ C \rightarrow 2\mathrm{P}^{C/V}$. 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 $2\mathrm{P}^{C/VP}$ 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 2*w*+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:

$$h_{d_i}(g \circ f(x,w,c)) = \left\{ \begin{array}{c} \left\langle g(x_{h_1}^{-w}),\ldots g(x_{h_1}^{-1}),(x,1),g(x_{h_1}^{+1}),\ldots g(x_{h_1}^{+w}) \right\rangle \\ \ldots\ldots\ldots. \\ \left\langle g(x_{h_m}^{-w}),\ldots g(x_{h_m}^{-1}),(x,|I_x|),g(x_{h_m}^{+1}),\ldots g(x_{h_m}^{+w}) \right\rangle \end{array} \right\} \quad (1)$$

In this paper, the problem is that the collocation of the central word is ordered according to word senses. Figure 1 show the overall process for this purpose.

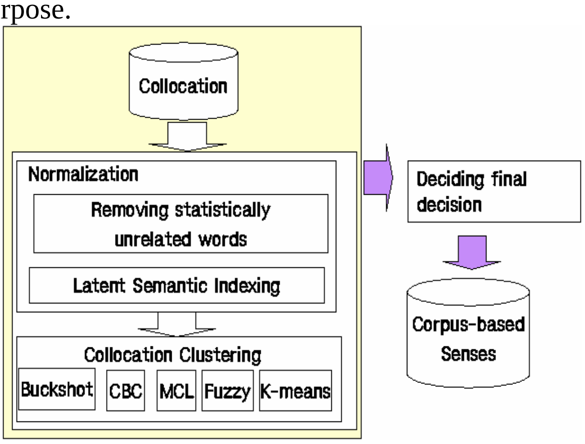

Figure 1 Processing for detecting sense boundary

3 Automatic clustering of collocation

For extracting practical senses, the contextual words for a central word are clustered by analyzing the pattern of the surrounding words. With this method, we can get the collocation without sense ambiguity, and also discover the practical sense boundary.

In order to extract the correct sense boundary from the clustering phase, it needs to remove the noise and trivial collocation. We call this process normalization, and it is specifically provided as [8]. The statistically unrelated words can be said that the words with high frequency appear regardless of their semantic features. After deciding the statistically unrelated words by calculating tf·idf values, we filtered them from the original surrounding words. The second normalization is using LSI (Latent Semantic Indexing). Throughout the LSI transformation, we can remove the dimension of the context vector and express the hidden features into the surface of the context vector.

3.1 Discovering sense boundary

We discovered the senses of the homonyms with clustering the normalized collocation. The clustering classifies the contextual words having similar context – the contextual words having similar pattern of surrounding words - into same cluster. Extracted clusters throughout the clustering symbolize the senses for the central words and their collocation. In order to extract clusters, we used several clustering algorithms. Followings are the used clustering methods:

- K-means clustering (K) (Ray and Turi, 1999)
- Buckshot (B) (Jensen, Beitzel, Pilotto, Goharian and Frieder, 2002)
- Committee based clustering (CBC) (Patrick and Lin, 2002)
- Markov clustering (M1, M2)[1] (Stijn van Dongen, 2000)
- Fuzzy clustering (F1, F2)[2] (Song, Cao and Bruza, 2003)

Used clustering methods cover both the popularity and the variety of the algorithms – soft and hard clustering and graph clustering etc. In all clustering methods, used similarity measure is the cosine similarity between two sense vectors for each contextual word.

We extracted clusters with these clustering methods, tried to compare their discovered senses and the manually distributed senses.

3.2 Deciding final sense boundary

After clustering the normalized collocation, we combined all clustering results and decided the optimal sense boundary for a central word.

$$h_{d_i}(g \circ f(x,w,c)) = S_{xd_i} = \{h_{1d_1}^{x},\ldots, h_{md_i}^{x}\} \quad (2)$$
$$(m = num(x, d_i))$$
$$D = \{d_0,\ldots, d_i,\ldots, d_n\}$$
$$S_x = \{s_{x0}, s_{x1},\ldots, x_{xm}\}$$

In equation (2), we define equation (1) as S_{xdi}, this means extracted sense boundary for a central word x with d_i. The elements of *D* are the applied clustering methods, and S_x is the final combination results of all clustering methods for *x*.

[1] M1and M2 have different translating methods between context and graph.

[2] F1and F2 are different methods deciding initial centers.

This paper proposes the voting of applied clustering methods when decides final sense boundary like equation (3).

$$Num(x) = \max_{d_i \in D} \{num(w, d_i)\} = S_x \qquad (3)$$

We determined the number of the final sense boundary for each central word with the number of clusters that the most clustering algorithms were extracted.

After deciding the final number of senses, we mapped clusters between clustering methods. By comparing the agreement, the pairs of the maximum agreement are looked upon the same clusters expressing the same sense, and agreement is calculated like equation (4), which is the agreement between *k*-th cluster with *i*-th clustering method and *l*-th cluster with *j*-th clustering method for central word x.

$$agreement = \frac{\{h^{x}_{kd_i}\} \cap \{h^{x}_{ldj}\}}{\{h^{x}_{kd_i}\} \cup \{h^{x}_{ldj}\}} \qquad (4)$$

$$Vot(S_x, w) = \sum_{x_k \in V} \max_{d_i \in D} \{h_{d_i}(g \circ f(x, w, c))\} \qquad (5)$$

$$\vec{z}_{S_x} = (\frac{1}{N_n}\sum w_{a_1}, \frac{1}{N_n}\sum w_{a_2}, \ldots, \frac{1}{N_n}\sum w_{a_n}) \qquad (6)$$

The final step is the assigning elements into the final clusters. In equation (5), all contextual words *w* are classified into the maximum results of clustering methods. New centers of each cluster are recalculated with the equation (6) based on the final clusters and their elements.

Figure 2 represents the clustering result for the central word ‘chair’. The pink box shows the central word ‘chair’ and the white boxes show the selected contextual words. The white and blue area means the each clusters separated by the clustering methods. The central word ‘chair’ finally makes two clusters. The one located in blue area contains the collocation for the sense about ‘the position of professor’. Another cluster in the white area is the cluster for the sense about ‘furniture’. The words in each cluster are the representative contextual words which similarity is included in ranking 10.

4 Experimental results

We extracted sense clusters with the proposed methods from the large-scaled corpus, and compared the results with the sense distribution of the existing thesaurus. Applied corpus for the experiments for English and Korean is Penn tree bank[3] corpus and KAIST[4] corpus.

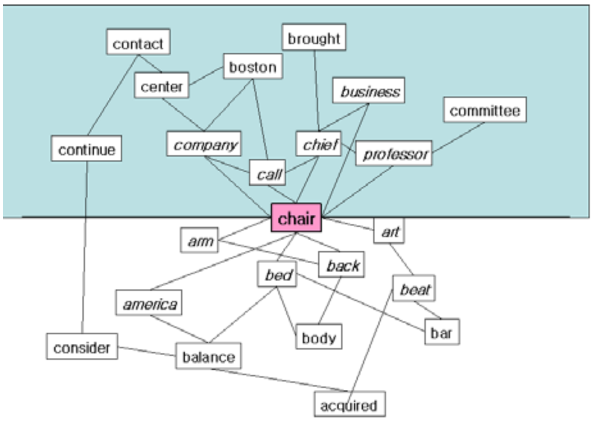

Figure 2 The clustering example for 'chair'

For evaluation, we try to compare clustering results and sense distribution of dictionary. In case of English, used dictionary is WordNet 1.7[5] - Fine-grained (WF) and coarse-grained distribution (WC). The coarse-grained senses in WordNet are adjusted sense based on corpus for SENSEVAL task. In order to evaluate the practical word sense disambiguation systems, the senses in the WordNet 1.7 are adjusted by the analyzing the appearing senses from the Semcor. For the evaluation of Korean we used Korean Unabridged Dictionary (KD) for fine-grained senses and Yonsei Dictionary (YD) for corpus-based senses.

Table 1 shows the clustering results by each clustering algorithms. The used central words are 786 target homonyms for the English lexical samples in SENSEVAL2[6]. The numbers in Table 1 shows the average number of clusters with each clustering method shown chapter 3 by the part of speech. WC and WF are the average number of senses by the part of speech.

In Table 1 and 2, the most clustering methods show the similar results. But, CBC extracts more clusters comparing other clustering methods. Except CBC other methods extract similar sense distribution with the Coarse-grained WordNet (WC).

	Nouns	Adjectives	Verbs	All
K	3	3.046	3.039	3.027
B	3.258	3.218	3.286	3.266
CBC	6.998	3.228	5.008	5.052
F1	3.917	2.294	3.645	3.515
F2	4.038	5.046	3.656	4.013
Final	3.141	3.08	3.114	3.13
WC	3.261	2.887	3.366	3.252
WF	8.935	8.603	9.422	9.129

Table 1 The results of English

[3] http://www.cis.upenn.edu/~treebank/home.html

[4] http://kibs.kaist.ac.kr

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

[6] http://www.cs.unt.edu/~rada/senseval/

	K	B	C	F1	F2	M1
Nouns	2.917	2.917	5.5	2.833	2.583	4.083
	KD	YD	M2			
Nouns	11.25	3.333	3.833			

Table 2 The results of Korean

Table 3 is the evaluating the correctness of the elements of cluster. Using the sense-tagged collocation from English test suit in SENSEVAL2[7], we calculated the average agreement for all central words by each clustering algorithms.

K	B	C	F1	F2
98.666	98.578	90.91	97.316	88.333

Table 3 The average agreement by clustering methods

As shown in Table 3, overall clustering methods record high agreement. Among the various clustering algorithms, the results of K-means and buckshot are higher than other algorithms. In the K-means and fuzzy clustering, the deciding random initial shows higher agreements. But, clustering time in hierarchical deciding is faster than random deciding

5 Conclusion

This paper proposes the method for boundary discovery of homonymous senses. In order to extract practical senses from corpus, we use the collocation from the large corpora and the clustering methods.

In these experiments, the results of the proposed methods are different from the fine-grained sense distribution - manually analyzed by the experts. But the results are similar to the coarse-grained results – corpus-based sense distribution. Therefore, these experimental results prove that we can extract practical sense distribution using the proposed methods.

For the conclusion, the proposed methods show the similar results with the corpus-based sense boundary.

For the future works, using this result, it'll be possible to combine these results with the practical thesaurus automatically. The proposed method can apply in the evaluation and tuning process for existing senses. So, if overall research is successfully processed, we can get a automatic mechanism about adjusting and constructing knowledge base like thesaurus which is practical and containing enough knowledge from corpus.

There are some related works about this research. Wortchartz is the collocation dictionary with the assumption that Collocation of a word expresses the meaning of the word (Heyer, Quasthoff and Wolff, 2001). (Patrick and Lin, 2002) tried to discover senses from the large-scaled corpus with CBC (Committee Based Clustering) algorithm.. In this paper, used context features are limited only 1,000 nouns by their frequency. (Hyungsuk, Ploux and Wehrli, 2003) tried to extract sense differences using clustering in the multi-lingual collocation.

6 Acknowledgements

This work has been supported by Ministry of Science and Technology in Korea. The result of this work is enhanced and distributed through Bank of Language Resources supported by grant No. R21-2003-000-10042-0 from Korea Science & Technology Foundation.

References

Ray S. and Turi R.H. 1999. *Determination of Number of Clusters in K-means Clustering and Application in Colour Image Segmentation*, In "The 4th International Conference on Advances in Pattern Recognition and Digital Techniques", Calcuta.

Heyer G., Quasthoff U. and Wolff C. 2001. *Information Extraction from Text Corpora*, In "IEEE Intelligent Systems and Their Applications", Volume 16, No. 2.

Patrick Pantel and Dekang Lin. 2002. Discovering Word Senses from Text, In "ACM Conference on Knowledge Discovery and Data Mining", pages 613–619, Edmonton.

Hyungsuk Ji, Sabine Ploux and Eric Wehrli. 2003, *Lexical Knowledge Representation with Contexonyms*, In "The 9th Machine Translation", pages 194-201, New Orleans

Eric C.Jensen, Steven M.Beitzel, Angelo J.Pilotto, Nazli Goharian, Ophir Frieder. 2002, *Parallelizing the Buckshot Algorithm for Efficient Document Clustering*, In "The 2002 ACM International Conference on Information and Knowledge Management, pages 04-09, McLean, Virginia, USA.

Stijn van Dongen. 2000, *A cluster algorithm for graphs*, In "Technical Report INS-R0010", National Research Institute for Mathematics and Computer Science in the Netherlands.

Song D., Cao G., and Bruza P.D. 2003, *Fuzzy K-means Clustering in Information Retrieval*, In "DSTC Technical Report".

Saim Shin and Key-Sun Choi. 2004, Automatic Word Sense Clustering using Collocation for Sense Adaptation, In "Global WordNet conference", pages 320-325, Brno, Czech.

[7] English lexical sample for the same central words