Language, Information and Computation (PACLIC12), 18-20 Feb, 1998, 236-247

Word-Sense Classification by Hierarchial Clustering

Ken Y.K. Lau and Robert W.P. Luk

Department of Computing, Hong Kong Polytechnic University
Email: csrluk@comp.polyu.edu.hk

Abstract

This paper investigates the use of clustering techniques in word-sense classification, which identifies different contexts that a word was used with the same or similar sense. For simplicity, we have used the hierarchical clustering techniques: single- and complete-linkage, and we showed that the latter is a more suitable technique from our performance measurements (i.e. recall and precision) compared with manually grouping different contexts of similar meaning. We found that the use of part-of-speech tags and fixed-length context has better clustering performance than without part-of-speech tags and sentence context, respectively. The differences between manually identified groups of different contexts are measured in terms of recall and precision at about 80%, which are not very different from the average recall and precision performance of complete-linkage clustering at 80% and 75%, respectively.

1. Introduction

Lexicographers can access, collect and analyze a large volume of language data by computer. The processed data are often put in the form of concordances as show in Figure 1. Lexicographers use these concordances to identify the different word senses and usually representative word senses are registered in the dictionary or glossary. However, identification of different word senses is a formidable task of an objective lexicographer because (s)he has to browse through $f(i)$ number of concordances if the word i has occurrence frequency $f(i)$. Since at most there are $f(i)$ different senses, the lexicographer has to perform at most $f(i)^2$ amount of comparison between concordances. Although computers cannot write down the senses of each word in the concordance, computers can assist the lexicographer by identifying the different groups of concordances for which the word has similar meaning. This would save the lexicographer from browsing through a large list of concordances and the lexicographer can select examples from groups of concordances of similar meaning, and register and write down those representative word senses.

..... 眼眶和耳根疼痛，并伴有高燒，嚴 [重] 的可導致死亡。委內瑞拉在去年１２ .....
..... 試航試吊救生艇時，有３人不慎受 [重] 傷，情況十分危急。駐滬海軍４３０ .....
..... 炮艇緊急起航，趕赴出事地點，將 [重] 傷員火速送往上海市區醫院搶救。羊 .....
..... 總產值占上海全市工業總產值的比 [重] ，由前年的百分之二上升到去年的百 .....
..... 度的增加。羊＃今年冶金工業要把 [重] 點放到增強行業發展后勁上來羊帶著 .....

Figure 1: A list of different concordances found in the PH corpus [1] for the character/word 重.

Apart from assisting lexicographers, automatically grouping different concordances of the word with similar meaning has found applications in document retrieval [2] since assigning words with different sense tags differentiates their meaning and therefore improve the precision performance of retrieval systems.

There are already many work reported in grouping different words based on their concordances for the automatic construction of thesauri [3,4]. However, there are relatively few work reported in grouping concordances with similar meaning, particularly for Chinese. Yarowsky [5] was advocating one sense per collocation (or concordance) so that word senses registered are only a representative or generic meaning of the words extracted from their different occurrences. Schutze [6] applied the vector model to written English contexts in order to measure similarity and in order to group concordances because the vector model has direct relevance to document retrieval. The performance is measured in terms of using the clustered senses to disambiguate a word in different contexts for effective retrieval. Park *et al.* [7] used genetic algorithm to automatically classify concordances or dictionary definitions of Korean words. The performance is measured by recall and precision compared with groupings of concordances derived manually. For Chinese, we are unaware of any report to group concordances of similar meaning of a word. For simplicity, we used the classical hierarchical clustering

techniques instead of the more sophisticated ones (e.g. genetic algorithm). At this preliminary stage, we focus on how different parameter settings of clustering affect performance.

2. Classification

More formally, let H_k denotes the set of concordances $\{h_{k,1}, h_{k,2}, \ldots, h_{k,m}\}$ of word k and H denotes all the concordances (i.e. $H = \cup_k H_k$). A concordance h is a triple (w, x, y) where w, y are strings over Σ^* and x is a string over Σ. A similarity measure is a function $s: H \times H \rightarrow \mathfrak{R}$ that maps the concordances into a scalar value where $\mathfrak{R}$ is the set of semi-definite positive real numbers. The word-sense classification problem is to find a function $\varphi: C_i \rightarrow T$ that maps a concordance to a sense tag t in a set T. Note that $|T| \leq |C_i|$ and therefore φ is not likely to be injective. Here, φ is found by hierarchical clustering technique. The basic idea is to combine concordances that are similar on the basis of the measure s. In each iteration, two clusters or sets of concordances are combined until the number the similarity between all the clusters of concordances fall below a given threshold. Then, a unique tag $t \in T$ is assigned to each cluster.

2.1 Similarity Measures

A similarity measure s describes the relationship between two individuals (or concordances in this case), given the values of a set of p variates common to both. In this case, the variates are identified as the characters in the concordances. Usually, the measure s is derived by counting the presence and absence of the variates between both individuals by some combination techniques. However, we hypothesize that the meaning of the keyword is mainly determined by the words which are closest to the keyword. Thus, the characters that have a shorter distance from the keyword is more important than those characters that are far away from the keyword.

For every two concordances, h_1 and h_2, the similarity between these two concordances is the sum of the similarity of each matching character C that occurs in h_1 and h_2. The number of characters between the matching character C and the keyword K is called the distance between C and K. The similarity value for each matched character C is geometrically weighted with a constant ratio R (where $0 < R < 1$) by the distance between the i^{th} character C_i and the keyword K in h_1 and the distance between character C_i in S_1 and character C_j in S_2, where $C_i = C = C_j$. Here, the suffixes i and j of C are the positions and therefore the distances from the keyword. Figure 2 shows the distances between the matched characters and the keyword and Figure 3 show how different geometric ratio R controls the weighting of the similarity values with respect to the distance between the matched character and the keyword.

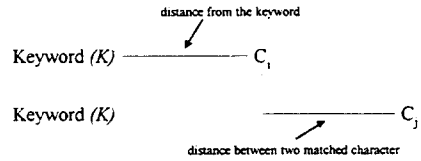

Figure 2. Calculation between two sentences

Consider the following two sentences, although they have the same character 理, their meaning is not the same.

.....領導正確性和科學性的一個 [重] 要 保 証，從而更加堅定了從事 理論教 .....
.....及近几年發展鋼鐵工業要著 [重] 抓好治理整頓和深化改革，全行業要 .....

We add another geometric weight to the similarity value. The weight has geometric ratio R and scaled by the number of characters between the two matching characters C_i and C_j (i.e. $|i - j|$ where $||$ returns the absolute value).

Hence, the similarity measure s between two concordances h_1 and h_2 is defined as:

$$s(h_1, h_2) = \sum_{(c,i)\in E(\Xi(h_1))} \sum_{(c,j)\in E(\Xi(h_2))} \max\{R^i \times R^{|i-j|}, 0.1\} + \sum_{(c,i)\in E(\Psi(h_1))} \sum_{(c,j)\in E(\Psi(h_2))} \max\{R^i \times R^{|i-j|}, 0.1\}$$

where $\Xi:(\Sigma^* x \Sigma x \Sigma^*) \rightarrow \Sigma^*$ and $\Psi:(\Sigma^* x \Sigma x \Sigma^*) \rightarrow \Sigma^*$ are the projection of the first and third elements, i.e. $\Xi((w,x,y)) = w$ and $\Psi((w,x,y)) = y$, respectively, Z is the set of integers and $E:\Sigma^* \rightarrow 2^{(\Sigma, Z)}$ returns the set of characters in the input and their corresponding leftmost position.

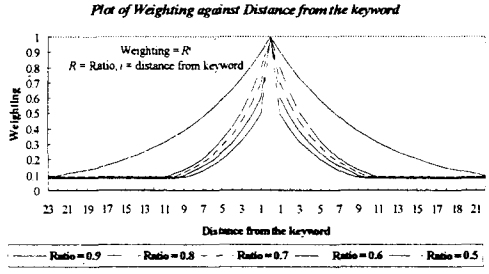

Figure 3. The relationship between the weighting and the distacnce from the keyword

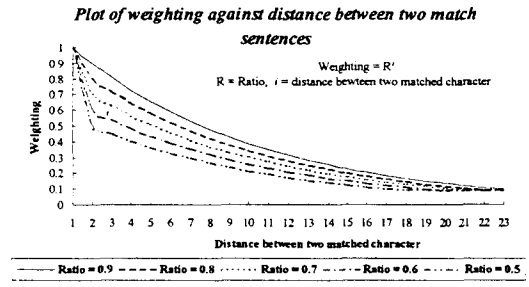

Figure 4. The relationship between two matched sentences and their distance

2.2. Linkage Techniques

Hierarchical clustering begins with the computation of a similarity or distance matrix between the entities, and end with a dendrogram showing the successive fusions of individuals, which culminates at the stage where all the individuals are in one group. At any particular stage, the methods fuse individuals or groups of individuals which are closest. Figure 5 demonstrate a typical hierarchical clustering technique.

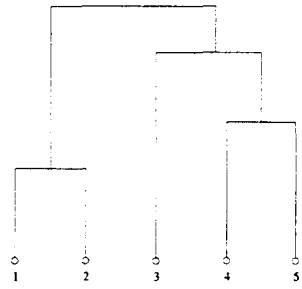

Figure 5. Dendrogram

2.2.1 Nearest Neighbor / Single Linkage Method

For single link clustering, groups are fused according to the similarity or distance between their nearest members i.e. the groups with the smaller distance or higher similarity are being fused. Each fusion joins two groups together and decreases the total number of groups by one. For this method, the distance / similarity between groups is defined as the distance / similarity between their closest members. The following example demonstrates the nearest neighbor clustering and Figure 6 shows the corresponding dendrogram.

$$D_1 = \begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ 1 & 0.0 & 2.0 & 6.0 & \boxed{10.0} & 9.0 \\ 2 & 2.0 & 0.0 & 5.0 & 9.0 & 8.0 \\ 3 & 6.0 & 5.0 & 0.0 & 4.0 & 5.0 \\ 4 & 10.0 & 9.0 & 4.0 & 0.0 & 3.0 \\ 5 & 9.0 & 8.0 & 5.0 & 3.0 & 0.0 \end{array}$$

$d_{(14)2} = \max\{d_{12}, d_{24}\} = d_{23} = 9.0$
$d_{(14)3} = \max\{d_{13}, d_{34}\} = d_{24} = 6.0$
$d_{(14)5} = \max\{d_{15}, d_{45}\} = d_{25} = 9.0$

$$D_2 = \begin{array}{c|cccc} & 1,4 & 2 & 3 & 5 \\ 1,4 & 0.0 & \boxed{9.0} & 6.0 & 9.0 \\ 2 & 2.0 & 0.0 & 5.0 & 8.0 \\ 3 & 6.0 & 5.0 & 0.0 & 5.0 \\ 5 & 9.0 & 8.0 & 5.0 & 0.0 \end{array}$$

$$d_{(14\,2)\,3} = \max\{d_{14\,3}, d_{23}\} = d_{143} = 6.0$$
$$d_{(14\,2)\,5} = \max\{d_{14\,5}, d_{25}\} = d_{145} = 9.0$$

$$D_3 = \begin{array}{c|ccc} & 1,4,2 & 3 & 5 \\ 1,4,2 & 0.0 & 6.0 & \boxed{9.0} \\ 3 & 6.0 & 0.0 & 5.0 \\ 5 & 9.0 & 5.0 & 0.0 \end{array}$$

$$d_{(142\,5)\,3} = \max\{d_{142\,3}, d_{5\,3}\} = d_{142\,3} = 6.0$$

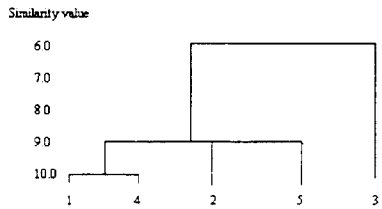

Figure 6: Nearest neighbor clustering of the example with a tie case.

2.2.2 Furthest Neighbor / Complete Linkage Method

For complete linkage clustering, it is very similar to single linkage. But the distance / similarity between two groups is defined as the distance between their remote pair of individuals. So the distance matrix D_1 of the previous will be as follow

$$D_1 = \begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ 1 & 0.0 & 2.0 & 6.0 & \boxed{10.0} & 9.0 \\ 2 & 2.0 & 0.0 & 5.0 & 9.0 & 8.0 \\ 3 & 6.0 & 5.0 & 0.0 & 4.0 & 5.0 \\ 4 & 10.0 & 9.0 & 4.0 & 0.0 & 3.0 \\ 5 & 9.0 & 8.0 & 5.0 & 3.0 & 0.0 \end{array} = 2.0$$

$$d_{(14)\,2} = \min\{d_{12}, d_{42}\} = d_{12} = 2.0$$
$$d_{(14)\,3} = \min\{d_{13}, d_{43}\} = d_{43} = 4.0,$$
$$d_{(14)\,5} = \min\{d_{14}, d_{45}\} = d_{45} = 3.0$$

By choosing the furthest neighbor on each fusion, the final result of the previous example is shown in Figure 7.

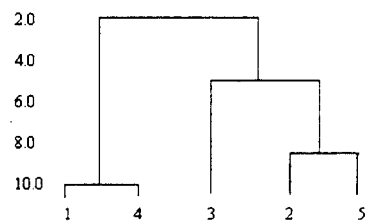

Figure 7: Complete Linkage Dendrogram

2.3 Thresholding

The vertical distance in the dendrogram represents the distance / similarity value for which the clusters at the same vertical level are fused. By defining a distance / similarity threshold, cluster formed immediately below the threshold are recognized as valid grouping of concordances. Clustered formed above the threshold are not considered.

For example, Figure 8 is the dendrogram of the keyword 重 in single linkage clustering. The number of meanings of the keyword 重 is determined by the threshold at certain vertical level in the dendrogram. At threshold level 0, there are two groups which represent keyword 重 have two meaning. At threshold level 1, 重 have three meanings, at threshold 2, 重 have four meanings and so on.

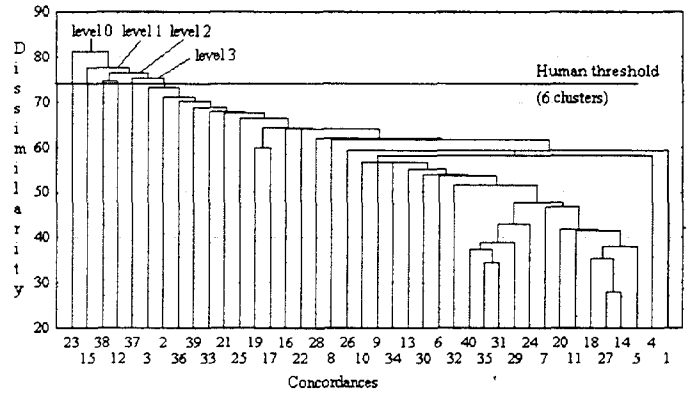

Figure 8: Single-linkage clustering of the keyword 重.. The Human threshold is derived from knowing the number of clusters that human identified and then search the equivalent threshold value that produces the same number of clusters as human (i.e. 6 clusters in this case).

3. Evaluation

Two corpora were used for evaluation. The first corpus, called PH, is from the National University of Singapore (NUS). The corpus has about 3 million Chinese characters, collected from Xin Hua News Agency (新華社) but there is no part-of-speech tagging. The other corpus with part-of-speech tagging is from Chinese Knowledge Information Processing Group (中文詞知識庫小組) under the Institute of Information Science Academia Sinica (中文研究院資訊科學研究所). The corpus contains more than 3 hundred thousand Chinese characters.

3.1 Performance Measurement

The performance measurement, called accuracy, is based on comparing automatically found (machine) clusters of concordances with those manually derived (i.e. human clusters). The accuracy is the product of two measurements called recall and precision. The recall is defined as the number of concordances matched (i.e. in both machine and human clusters) over the total number of concordances in the human cluster and the precision is defined as the number of concordances matched over the total number of concordances in the machine cluster. Although we can compute the recall and precision between every machine cluster and every human cluster, it is obvious that the recall and precision of clusters that are not intended to be identical are low. We need to find which machine cluster should correspond to which human cluster and then we compute the recall and precision for these matching pairs of clusters rather than for all the machine and human clusters.

We can visualize the problem of matching machine and human clusters as finding a match in a bipartite graph $G(M \cup S, E)$ where M is the set of vertices representing each machine cluster, S is the set of vertices representing each human cluster and $E \subseteq M \times S \times \Re$ is the set of weighted edges. The weight $w_{i,j}$ of an edge $(MV_i, HV_j, w_{i,j})$ is the accuracy (i.e. recall times precision) of the machine cluster MV_i with respect to the human cluster HV_j. The matching problem can be considered as finding a minimum spanning tree τ that connects all the vertices in M with S and that optimizes the matching score defined as the negative sum of all the weights of the edges in τ, i.e.:

$$MatchScore = -1 \times \sum_{(MV_i, HV_j, w_{i,j}) \in T} w_{i,j}$$

We have used the Kruskal [9] algorithm to find the minimum spanning tree instead of Prime [10] because the former can find a forest instead of a tree. Figure 9 show a spanning tree (or forest) for keyword 重.

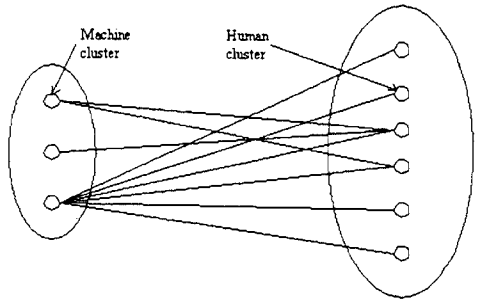

Figure 9: The minimum spanning tree of the keyword 重 with 3 machine clusters compared with 6 human clusters.

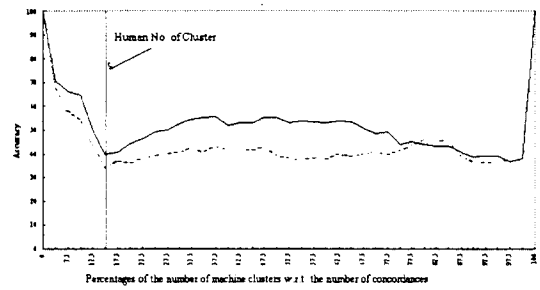

Figure 10: Accuracy performance of the keyword for different numbers of machine cluster. The solid line indicates the performance with part-of-speech tagging and the dotted line without part-of-speech tagging.

In our clustering technique, we can construct different number of grouping with different level of threshold. So with different number of machine clusters, we can plot a graph showing the accuracy for the different number of machine clusters. Figure 10 shows the performance of our clustering in classifying 40 concordances with the keyword 重. The filled line represent the result that included all the part-of-speech tags in the concordances, and the dotted line represent the result that excluded the part-of-speech tags in the same concordances. From the result in Figure 19, the curve that has part-of-speech obtained a better result than the curve without part-of-speech. Also the curve in Figure 19 shows that the accuracy is about 40 % when the machine divided the sentences into 6 groups which is the same number of human clusters. When the number of machine cluster is smaller than the human grouping, the human cluster will join together to form a larger group to match with the machine cluster. Thus, the accuracy will be higher. However, the actual clustering may not be desirable because when two human clusters join together to form a new cluster, one of the meaning from the original cluster will be lost. This is the case where the machine is under classifying the meaning of the word. Figure 11 demonstrate the case of machine under classifying the meaning of the word.

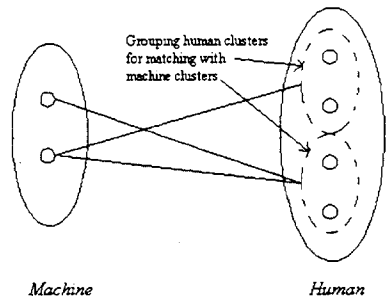

Figure 11: Under-classification of the meaning of a word.

On the other hand, when the number of machine cluster is greater than the number of human cluster, the machine cluster will join together to form a larger group to match with human clusters. Figure 12 show the case where the machine grouping is over classifying.

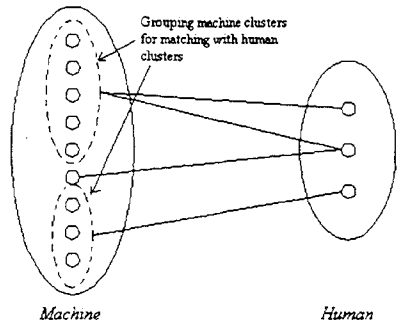

Figure 12: Over-classification of the meaning of a word.

For grouping a number of small clusters into a larger cluster due to the over classification of the keyword, this may not actually reflect the case where some of the meaning is lost. Two reasons for this phenomenon are:

- Some of the meaning of the keyword is very similar, such as the keyword 重 has the meaning of *repeat* (複) when the word is an adjective and has another of *again* (再) when the word is an adverb. So sometimes human may groups these two meaning into the same group;
- Our clustering methods will form a binary tree. So it may separate words with similar meaning into different groups. The joining of small groups into a larger cluster will join the words with similar meaning into the same group.

Thus, the curve rises smoothly when the number of machine cluster is greater than the human grouping (Figure 10). However sometimes when the meaning of the word is very clear, the over classification of the machine cluster may reflect the grouping of different meaning for the word into the same group. Figure 13 shows another result of classifying 40 concordances with the keyword 由. For the keyword 由, it only contains three meanings from the dictionary, (1) *reason (原因)* for noun, (2) *from (從)* for verb and (3) *free (隨便)* and their meaning is distinct. So after the number of machine cluster is greater than the number of human grouping, the curve rises rapidly and this is the result of grouping different meaning of the word into the same group.

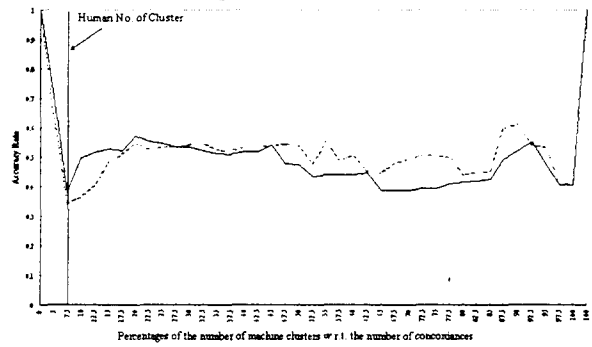

Figure 13: Accuracy of clustering the concordances of the keyword 由 for different numbers of machine cluster. The solid and dotted lines represent accuracy with and without part-of-speech tagging, respectively.

We can examine in more details about under- and over-classification by plotting the recall and precision variation against different number of machine clusters. Figure 14 and 15 show the recall and precision of the keyword 重 for different number of machine clusters. Figure 16 and 17 show the recall and precision of the keyword 由.

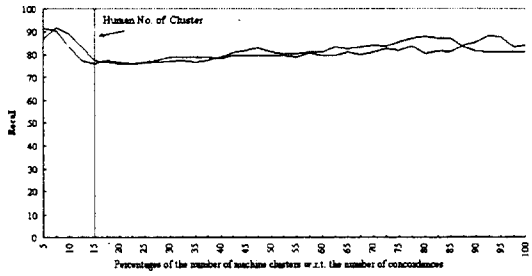

Figure 14: Recall of clustering the concordances of the keyword 重 for different numbers of machine cluster. The solid and dotted lines represent accuracy with and without part-of-speech tagging, respectively.

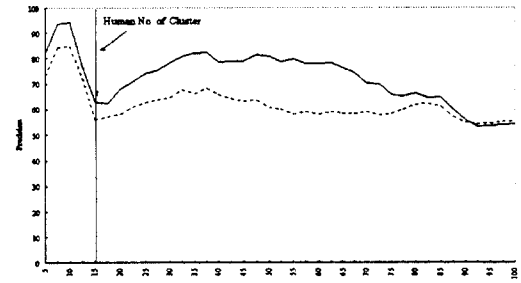

Figure 15: Precision of clustering the concordances of the keyword 重 for different numbers of machine cluster. The solid and dotted lines represent accuracy with and without part-of-speech tagging, respectively.

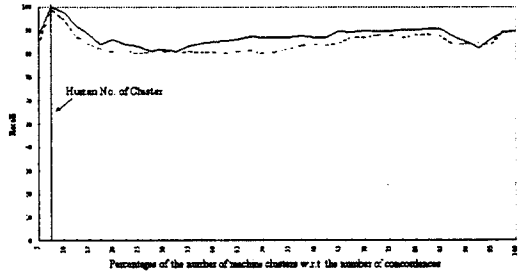

Figure 16: Recall of clustering the concordances of the keyword 由 for different numbers of machine cluster. The solid and dotted lines represent accuracy with and without part-of-speech tagging, respectively.

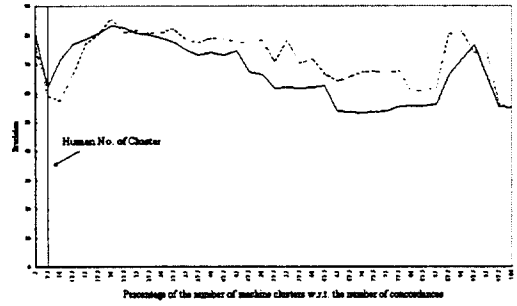

Figure 17: Precision of clustering the concordances of 由 for different numbers of machine cluster. The solid and dotted lines represent accuracy with and without part-of-speech tagging, respectively.

When the number of human cluster is greater than the number of machine cluster, the recall is much higher than the precision. This is because a larger group is formed in under-classification, which leads to better recall. When the number of machine cluster is greater than the number of human group, this will lead to a rise of the precision because over-classification joins the human clusters to form one larger group. From Figure 17, the precision rises rapidly due to incorrect grouping of machine cluster to form a larger group. A different phenomenon occurs in Figure 15, the precision remains the same and rise smoothly when the number of machine cluster reduces. That means 重 can be classified, less dependent on the value of the threshold, which is desirable.

For our overall performance, we have selected 10 characters and we randomly selected 40 concordances of these characters from the corpus. Then, these concordances are classified manually. In order to have a general classification from the human clusters, the concordances are classified by 3 subjects because different person may have different grouping since they may have different interpretation of the meaning of the keyword in different concordances.

3.2 Keyword Selection

We compiled the number of occurrences of each character in the corpus, and then based on its occurrence, sort them in descending order. Next, we assigned each character with its part-of-speech tags. Based on the number of part-of-speech tag of each character, we sort them in descending order. We can extract those characters which are frequently occurring in the corpus and which contains as much part-of-speech tags as possible. In this way, we have a set of keywords that have many differentiated meaning as well as enough occurrences for evaluation.

No.	Keyword	Part-of-Speech from Dictionary (辭淵)	Part-of-Speech from corpus
1.	重	名詞、動詞、副詞、形容詞、形名詞	名詞、動詞、副詞
2.	由	名詞、動詞、置詞	名詞、動詞、介詞、連接詞
3.	盡	名詞、動詞	名詞、動詞
4.	結	名詞、動詞	名詞、動詞、副詞
5.	若	副詞、形容詞、連接詞	名詞、動詞、連接詞
6.	刻	名詞、動詞、副詞、形容詞	名詞、動詞、副詞
7.	惡	動詞、形容詞、助詞	名詞、動詞、副詞、形容詞
8.	隨	名詞、動詞、形容詞	動詞、副詞、介詞
9.	小	名詞、動詞、副詞、形容詞	名詞、動詞、副詞、形容詞
10.	效	名詞、動詞	名詞、動詞

Table 1. Sample Data and their part of speech

Table 1 shows the selected 10 keywords and their possible part-of-speech. Since their occurrences in the corpus are very high, we randomly selected a portion of sentences for human classification. Most of these sentences are given to three people to classify the word sense manually. So totally, there are 30 sample data to verify our system performance.

After selecting the keyword, we extract the concordance with the given keyword from the corpus with a fixed length context. The length of the context is set to 50 characters not including punctuation.

3.4 Single vs Complete Linkage

Comparing with human clusters of the same data in Figure 18, human identified six meanings for the keyword 重 and for threshold level 6, the size between human cluster and single linkage cluster differs significantly. For single linkage, the clustering between groups is performed for closest cluster or concordance. This will easily lead to successive clustering of concordances to the largest group giving a skewed dendogram. This is called kerning which is in general not desirable. Figure 19 and figure 20 demonstrate another example of single linkage cluster and human cluster. The size of each group between single linkage and human linkage are also quite different.

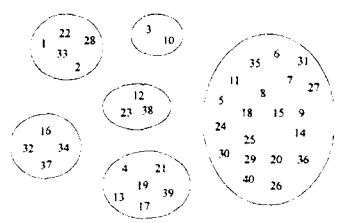

Figure 18: Comparison of human and machine clusters. Each number is the label t of a unique machine cluster. The number represents a particular concordance of the keyword 重 is classified to the corresponding machine cluster.

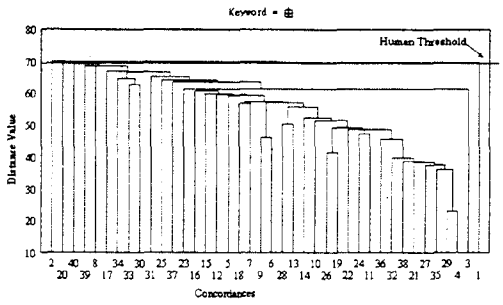

Figure 19: Clustering of the keyword 由 based on single-linkage. A small degree of skewness can be found towards the left where single concordances are grouped with the larger group of concordances.

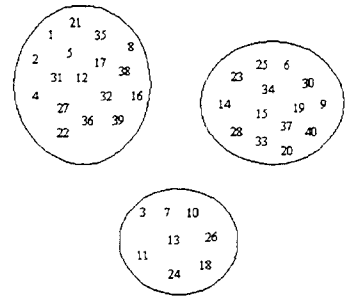

Figure 20: Comparison of human and machine clusters. Each number is the label t of a unique machine cluster. The number represents a particular concordance of the keyword 由 is classified to the corresponding machine cluster.

For complete linkage, since the distance between groups is defined as the distance between their most remote pair of individuals, this will force different cluster to form separately at lower level and will join together at upper level. Some clustering results are shown in Figure 21 and Figure 22. We can identify several clusters, which may be representing different meaning of the keyword that have been used in the concordances. Table 2 shows the comparison of the recall and precision performance between single- and complete-linkage. We showed that the complete linkage is better than the single linkage clustering. In particular, the precision of single-linkage is much lower than complete linkage, indicating that the effect of kerning is not desirable. Therefore, for the rest of our project, we will use complete linkage for our clustering technique.

Clustering Technique	Recall	Precision
Single Linkage	66.7	13.4
Complete Linkage	71.8	75.7

Table 2:A comparison of Single Linkage and Complete Linkage with the keyword 由

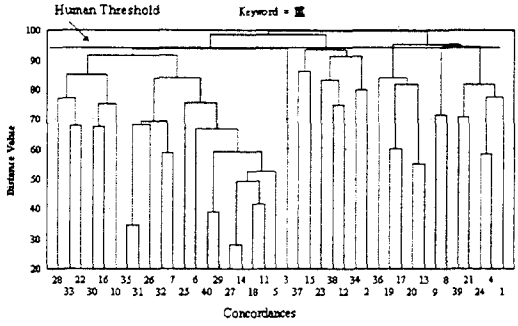

Figure 21: Clustering of the keyword 重 based on complete-linkage. Small groups are formed before larger groups.

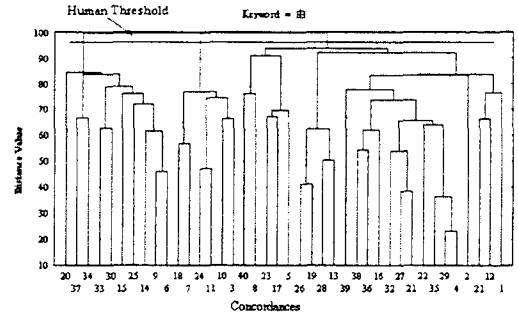

Figure 22: Clustering of the keyword 由 based on complete-linkage. Small groups are formed before larger groups.

3.5 Effects of Part-of-Speech Tagging

Figure 23 shows that the accuracy is higher for concordances with part-of-speech over those concordances without part-of-speech for most of the different number of machine clusters. We have concluded that corpus with part of speech can yield a better accuracy on word sense classification because:

1. the part-of-speech tags differentiate the same words with different meanings so that the similarity between the concordances of the keyword can be measured more accurately;
2. certain function words which play no part in determining the similarity value can be deleted so that chances of spurious association between concordances are reduced.

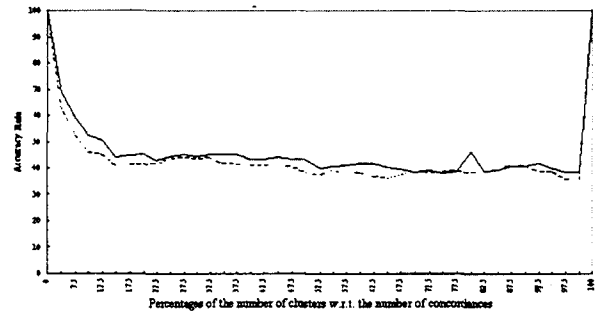

Figure 23: Consistent better accuracy with part-of-speech tagging (solid line) compared with without part-of-speech tagging (dotted line) for different numbers of machine cluster.

3.6 Sentence or Fixed-Length Context

For concordances retrieval, we extract complete sentences which are variable length and the location of the keyword also varies. For fixed length contexts, we extract 50 characters away from the keyword. Usually, it contains more than one sentences. After our similarity measures and clustering, the result is showed in Figure 24.

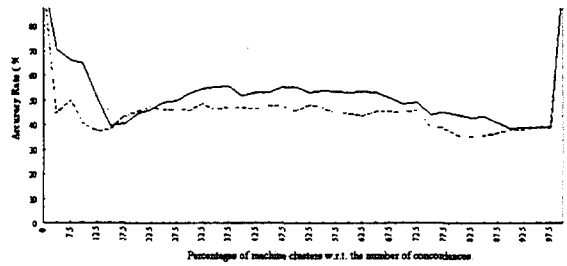

Figure 24 Consistent better accuracy with fixed-length contexts (solid line) compared with sentence contexts (dotted line) for different numbers of machine cluster.

In Figure 24, the accuracy of the fixed length contexts is higher than the complete sentences nearly for all percentages of clusters. It shows that for word sense classification, a fixed length contexts is better than complete sentences. This is because complete sentences can be very short with little information about the meaning of the keyword. For example, consider the following sentence with the keyword 走: *我走了。*.

Based on the contexts 我 and 了, it is difficult to classify the meaning of the word 走 between running (跑) and departing (離開).If the keyword is located at the beginning of the sentence or at the end of the sentence, there is no upper contexts or lower contexts to determine the meaning. On the other hand, fixed-length contexts ensure that there must have upper and lower contexts to determine the meaning of the keyword. Also, the fixed length contexts usually contain more than one sentences which would provide more information to determine the meaning of the keyword. Our finding is in accord with the practice of lexicographers [11].

4. Summary

We have shown that hierarchical clustering techniques can be applied to word-sense classification with comparable results to the agreement between manual word-sense classification. We showed that it is preferable to use fixed-length context over sentence-context, as well as the use of complete-linkage over single-linkage. We showed the variation of performance with different threshold values for defining different number of clusters. Typically, the shape of the curve is a bath tub, indicating that finer and finer classifications (i.e. smaller and smaller groups) have similar accuracy. This is desirable since the performance does not depend on the threshold value (for the non-extrema cases). The use of part-of-speech demonstrate a consistent improvement in accuracy of clustering over different number of machine clusters.

Acknowledgments

We would like to thank ROCLING, and Guo and Liu for providing the ROCLING balanced corpus and the PH corpus, respectively.

References

1. Guo, J. and H.C. Liu (1992) A Chinese corpus for pinyin-hanzi transcription, ISS Technical Report TR93-112-0, Institute of Systems Science, National University of Singapore.

2. Schutze, H. and J.O. Pedersen (1995) Information retrieval based on word senses, Proceedings of DAIR 95.

3. Brown, P.F. (1992) Class-based n-gram models of natural language, Computational Linguistics, 18:4.

4. Ushioda, A. (1996) Hierarchical clustering of words and application to NLP tasks, Proc. Of the Fourth Workshop on Very Large Corpora, Copenhagen, pp.28-41.

5. Yarowsky, D. (1993) One sense per collocation, Proceedings of ARPA Human Language Technology Workshop.

6. Schutze, H. (1992) Dimensions of meaning, Proceedings of Supercomputing 92,

7. Park, Y.J., S.K. Chung, M.S. Song (1997) Automatic classification of word senses using a genetic algorithm, Proceedings of ICCPOL 97, Vol II, pp. 426-427, April, Hong Kong.

8. Chen, K.J. and C-R. Huang (eds.) (1993) Chinese word class analysis, Technical Report 93-05, Chinese Knowledge Information Processing Group, Institute of Information Science, Academia Sinica, Taiwan.

9. Kruskal, J.B. Jr. (1956) On the shortest subtree of a graph and the traveling salesman problem, Proc. Of the American Mathematical Society, 7(1), 48-50.

10. Prim, R.C. (1957) Shortest connection networks and some generalizations, Bell System Technical Journal, 36, 1389-1401.

11. Hawks, P (1996) Contextual Dependency and lexical Sets, International Journal of Corpus Linguistics, vol.(1) 75-98.