A Clustering Algorithm for Chinese Adjectives and Nouns[1]

Yang Wen[1], Chunfa Yuan[1], Changning Huang[2]
[1]State Key Laboratory of Intelligent Technology and System
Deptartment of Computer Science & Technology, Tsinghua University,
Beijing 100084, P.R.C.
[2]Microsoft Research, China
Email: ycf@s1000e.cs.tsinghua.edu.cn

Key Words:

bidirectional hierarchical clustering, collocations, minimum description length, collocational degree, revisional distance

Abstract

This paper proposes a bidirctional hierarchical clustering algorithm for simultaneously clustering words of different parts of speech based on collocations. The algorithm is composed of cycles of two kinds of alternate clustering processes. We construct an objective function based on Minimum Description Length. To partly solve the problem caused by sparse data two concepts of collocational degree and revisional distance are presented.

1 Introduction

Recently research on the compositional frames (classification and collocational relationship of words) for Chinese words has been described in Ji et al. (1996)[1], Ji (1997)[2]. The objective of their work is to obtain the clusters of words of different parts of speech and to derive the collocational relationship between different clusters from the collocational relationship between words of different categories.

There are two ways to construct the clusters: One is to get clusters from thesaurus classified manually by linguists. But the fact is that words with the same meanings do not always have the same ability of collocating with other words. The method isn't fit for the NLP problems under our consideration. Another way is to get clusters automatically by computing on the distribution environments of words based on statistical method. The distribution environment of a word is the set of words of other parts of speech that can be collocated with it. We employ the second method in our work.

Previous research usually gets the clusters of words of a certain part of speech based on their distribution environments. But

[1] Supported by National Natural Science Foundation of China (69773031) and "973" Project (G1998030507).

we accept the assumption that the clustering processes of words of different parts of speech are inherently related. For example, having collocations between Chinese adjectives and nouns and if we take on nouns as entities and adjectives as features of nouns' distribution environments, we can obtain clusters of nouns and vice versa. The key of the relationship of the two clustering processes is that they use the same collocations. Therefore we consider the question of clustering the nouns and adjectives simultaneously. Li's work shows that they optimize the clustering results based on this viewpoint (Li et al., 1997)[3]. But they don't explain how to get initial clusters and their scale of problem is too small.

In this paper, we propose an algorithm named bidirectional hierarchical clustering to attempt answering the question.

2 Concepts

2.1 Problem Description

Our problem can be described as follows: given the set of adjectives A, the set of nouns N and the collocation instances, our system will construct a partition P_N over N and a partition P_A over A that respectively contain sets of nouns and sets of adjectives. And both partitions meet the condition that words in the same set (called cluster) have similar semantic distribution environment.

2.2 Partitions and Clusters

Let S be a set, $S_i \subset S(i=1,2,\cdots,n)$. If P_S={ S_i } satisfies that

$$(1)\bigcup_{i=1}^{n} S_i = S$$

$$(2)\ S_i \cap S_j = \varnothing, \forall i,j=1,2,\cdots n, i \neq j$$

Then P_S is a partition over S.

In this paper, we call $A_i \in P_A$ an "adjective cluster" and $N_i \in P_N$ a "noun cluster". And we want to obtain the composition of partitions $<P_A, P_N>$ as the clustering result.

2.3 Distance between Clusters

In order to measure the distance between clusters of the same part of speech, we use the following equations:

$$dis_A(A_i, A_j) = 1 - \left|\frac{\Phi_i \cap \Phi_j}{\Phi_i \cup \Phi_j}\right| \quad (1)$$

and

$$dis_N(N_i, N_j) = 1 - \left|\frac{\Psi_i \cap \Psi_j}{\Psi_i \cup \Psi_j}\right| \quad (2)$$

where Φ_i is the distribution environment of A_i and is make up of nouns which can be collocated with A_i. Ψ_i is the distribution environment of N_i and is composed of adjectives which can be collocated with N_i. Φ_j and Ψ_j follow similar definitions. This distance is a kind of Euclidean distance.

2.4 Collocational Degree

Since redundant collocations might be created during clustering, the concept "collocational degree" is used to measure the collocational relationship between a cluster and its distribution environment. The collocational degree is defined as the ratio of the existing collocation instances between the cluster and its distribution environment to all possible collocations generated by them. Thus,

$$\deg A_i = \frac{|\{a\phi \mid a \in A_i, \phi \in \Phi_i, a\phi \in C\}|}{|A_i||\Phi_i|} \quad (3)$$

and

$$\deg N_i = \frac{|\{n\psi \mid n \in N_i, \psi \in \Psi_i, n\psi \in C\}|}{|N_i||\Psi_i|} \quad (4)$$

where C is the set of all existing instances.

2.5 Redundant Ratio

After we get the collocational degree of a cluster, redundant ratio (marked as r) is calculated to measure the whole performance of the clustering result. We define the redundant ratio as 1 minus the ratio of all existing instances to all possible collocations generated by all clusters (including nouns and adjectives) and their distribution environments. So r is calculated as

$$r = 1 - \frac{2|C|}{\sum_i |A_i||\Phi_i| + \sum_i |N_i||\Psi_i|} \quad (5)$$

3 A Bidirectional Hierarchical Clustering Algorithm

Usually a hierarchical clustering algorithm [7] constructs a clustering "tree" by combining small clusters into large ones or dividing large clusters into small ones. The bidirectional hierarchical clustering algorithm proposed by us is composed of two kinds of alternate clustering processes.

The algorithm flow is described as follows:

1) Initially, regard every noun and adjective each as a cluster. Calculate the distances between clusters of the same part of speech.
2) Suppose without loss of generality that we choose to cluster nouns first. Select two noun clusters N_i & N_j of the minimum distance and integrate them into a new one N_i'.
3) Calculate the collocational degree of the new cluster. Adjust the sequence numbers of the original clusters and the relational information of adjective clusters.
4) Calculate the distances between the new cluster and other clusters.
5) Repeat from step 2) to 4) until the satisfaction of certain condition. For example, the number of the clusters has decreased to certain amount.[2]
6) Similarly, we can follow the same steps from 2) to 5) for constructing adjective clusters, completing one cycle of clustering processes of nouns and adjectives.
7) Repeat from step 2) to 6) until the

[2] In this paper, we set the proportion is 20%.

objective function[3] reaches the minimum value.

One advantage of this algorithm is that: when two clusters of nouns have similar distribution environments, they might be classified into one cluster. This information can be delivered to the clusters of adjectives that respectively collocate with them by the clustering process of nouns. Thus these clusters of adjectives have great possibility to be combined into one cluster, while the ordinary hierarchical clustering algorithm can not do it.

4 An Objective Function Based on MDL

The objective function is designed to control the processes of clustering words based on the Minimum Description Length (MDL) principle. According to MDL, the best probability model for a given set of data is a model that uses the shortest code length for encoding the model itself and the given data relative to it [4] [5]. We regard the clusters as the model for the collocations of adjectives and nouns. The objective function is defined as the sum of the code length for the model ("model description length") and that for the data ("data description length"). When the clustering result minimises the objective function, the bidirectional processes should be stopped and the result is the best probable one. The objective function based on MDL trade-offs between the simplicity of a model and its accuracy in fitting to the data, which are respectively quantified by the model description length and the data description length.

The following are the formulas to calculate the objective function L:

$$L = L_{mod} + L_{dat} \qquad (6)$$

L_{mod} is the model description length calculated as

$$\begin{aligned} &L_{mod} \\ &= -\sum_{i=1}^{k_A} \frac{1}{k_A} \log_2 \frac{1}{k_A} - \sum_{i=1}^{k_N} \frac{1}{k_N} \log_2 \frac{1}{k_N} + 1 \qquad (7) \\ &= \log_2(k_A k_N) + 1 \end{aligned}$$

Where k_A and k_N respectively denote the number of clusters of adjectives and nouns. "+1" means that the algorithm needs one bit to indicate whether the collocational relationship between the two clusters exists.

L_{dat} is composed of the data description length of adjectives and that of nouns, namely

$$L_{dat} = L_{dat}(A) + L_{dat}(N)$$

(8)

And the two types of data description length are calculated as follows

$$L_{dat}(A) = -\sum_{i=1}^{k_A} \frac{1}{k_A k_N} \sum_{j=1}^{|\Phi_i|} \log_2 \frac{1}{|A_i||N_k|} \qquad (9)$$

$$\forall j, \phi_j \in \Phi_i \text{ and } \phi_j \in N_k$$

$$L_{dat}(N) = -\sum_{i=1}^{k_N} \frac{1}{k_A k_N} \sum_{j=1}^{|\Psi_i|} \log_2 \frac{1}{|N_i||A_k|}$$

(10)

$$\forall j, \psi_j \in \Psi_i \text{ and } \psi_j \in A_k$$

5 Our Experiment

We take the words and collocations

[3] Described later in section 4.

gathered in Ni's Thesaurus [6] to test our algorithm. From Ni's thesaurus, we obtain 2,569 adjectives, 4,536 nouns and 37,346 collocations between adjectives and nouns.

Table 1 shows results of using 5 different revisional distance formulas discussed in the next section. Because the length of this paper is limited, we only give some examples (10 clusters for each part of speech) of clusters in section 8. We can see that the redundant ratio obviously decreases by using the revisional distance, and the result that has the lowest redundant ratio corresponds of the minimum value of the objective function. By human evaluation, most clusters contain the words that have similar meanings and distribution environments. So our algorithm proves to be effective for word clustering based on collocations.

Table 1: Results of different revisional distances

Revisional distance	k_A	k_N	L	r
Not used	409	550	20. 067	99. 01%
$dis' = -\overline{\deg} \ln dis$	397	610	20. 082	86. 96%
$dis' = dis / \overline{\deg}$	383	595	20. 002	78. 78%
$dis' = dis / \sqrt{\overline{\deg}}$	373	586	20. 017	80. 39%
$dis' = -dis \ln \overline{\deg}$	395	557	20. 007	80. 08%

However, the redundant ratio is still very large. The main cause is that existing instances are too sparse, covering only 0.32% of all possible collocations. So another advantage of our algorithm is that we can acquire many new reasonable collocations not gathered in the thesaurus. If we add the new collocations into initial thesaurus and execute the algorithm on new data set, the performance will have great potential to improve. It is further work that can be carried out in the future.

6 Discussions

6.1 Rivisional Distance

When we combine clusters into a new cluster, their distribution environments will be combined as well. The combination of clusters and their distribution environments might very likely generate redundant collocations that are not listed in the thesaurus. With the word clustering processes going on, there might be more and more redundant collocations. They will obviously affect the accuracy of the distances between clusters. When calculating the distances, the redundant collocations must be considered. So the question is how to revise the distance equation. Notice that the collocational degree defined in the above measures the collocational relationship between a cluster and its distribution environment. Obviously under the same distance, clusters having higher collocational degree have more higher similarity between each other (because they have more actual collocations) than those having lower collocational degree. So the collocational degree can be used to revise the distance equations.

There are two problems that should be considered when we design the revisional distance equations. The first one is to convert the collocational degrees of two clusters into one collocational degree as the revisional factor for distance equations. It is the average collocational degree, marked as $\overline{\deg}$, calculated by

$$\overline{\deg}_A = \frac{\deg A_i|\Phi_i||A_i| + \deg A_j|\Phi_j||A_j|}{(|A_i| + |A_j|)|\Phi_i \cup \Phi_j|} \quad (11)$$

and

$$\overline{\deg}_N = \frac{\deg N_i|\Psi_i||N_i| + \deg N_j|\Psi_j||N_j|}{(|N_i| + |N_j|)|\Psi_i \cup \Psi_j|} \quad (12)$$

In fact it is the collocational degree of the new cluster into which if we assume combining the two original clusters.

The second problem is that the revisonal distance equations should keep coherent of monotonicity with the original distance. It means that under the same average collocational degree, the revional distance should keep the same (or opposite) monotonicity with the original distance, and under the same original distance, the revional distance should keep the same (or opposite) monotonicity with the average collocational degree.

In this paper, four simple revisional distance equations are presented based on consideration of the upper two problems. They are:

a) $dis' = -\overline{\deg} \ln dis$

b) $dis' = \frac{dis}{\overline{\deg}}$

c) $dis' = \frac{dis}{\sqrt{\overline{\deg}}}$

d) $dis' = -dis \ln \overline{\deg}$

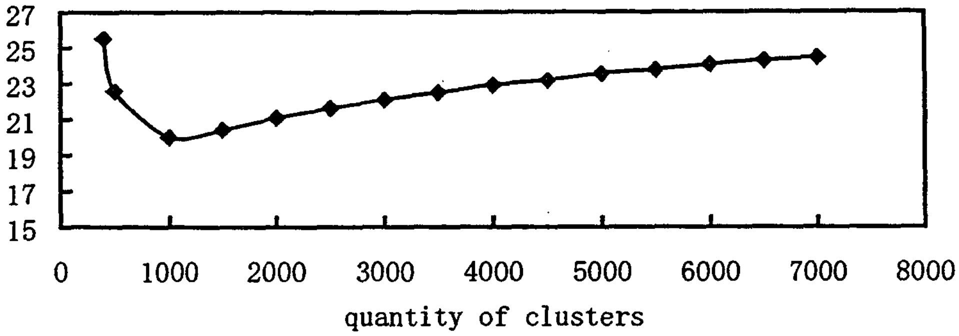
Where dis' denotes the revional distance and dis denotes the original distance.

From the comparison of the upper different results (shown in Table 1), we can draw the conclusion that using revisonal distance equations can increase the clustering accuracy remarkably.

6.2 Determinant of Objective Function's Minimum Value

The clustering algorithm terminates when the objective function is minimized. As a result it is very important to find out the function's minimum value. After analyzing the objective function, we find that it normally monotonically declines with clustering processes going on until it gets minimized. At the beginning, there are a large number of clusters with only one element in each of them. So the model description length is quite large while the data description length is quite small. Because the clustering process is hierarchical, every time when the combination occurs the number of clusters will decrease by one with the model description length's decreasing as well. At the same time the number of a certain cluster's elements will increase by one with the data description length's increment as well. However, the decrement is larger than the increment and it is getting smaller while the increment is getting larger. In this way, the objective function declines until the objective function reach its minimum value. If we continue to execute the algorithm, we will see that the value of the objective function rises very fast like as is shown in Figure 1.

Figure 1: Values of the Objective Functions



Therefore we choose a fairly simple way to avoid the appearance of the local optimum: When there are two consecutive increases in the objective function during one clustering process, stop the process and start another one. When two consecutive clustering processes are stopped due to the same reason, we assume that we have got the minimum value and stop the whole clustering process. In our future work we will try to find a better way to determine the minimum value of the objective function.

7 Conclusion & Future Work

In this paper we have presented a bidirctional hierarchical clustering algorithm of simultaneously clustering Chinese adjectives and nouns based on their collocations. Our preliminary experiments show that it can distinguish different words by their distribution environments. In addition, the clustering algorithm may help to find new collocations that are not in the thesaurus. This algorithm can also be extended to other collocation models, such as verb-noun collocations.

Our future work includes:

1) Because the sparsity of collocations is a main factor of affecting the word clustering accuracy, we can use the clustering results to discover new data and enrich the thesaurus.
2) As there are yet no adjustments to the hierarchical clustering results, we are considering using some iterative algorithm, such as K-means algorithm, to optimise the clustering results.

8 Attachment (Examples)

We give 10 clusters of each part of speech clustered by our algorithm (using revisional distance formula b) as follows:

8.1 Chinese Adjective clusters (10 of 383)

A1 风趣 天生 雄辩

A2 慷慨 荒唐 虚伪 乐观 糊涂 天真 冷静 单纯 缓和 荒谬 苦闷 悲观 烦恼 灰心 慎重 果断 苦恼 清醒 稚嫩 稚气 清高 迫切

A3 空虚 纯洁 脆弱 忧郁 消遥 充实 狂热

A4 糟糕 出色 优秀 杰出 卓越 合格 逊色 低劣 香

A5 卑劣 庸俗 卑鄙 下流 无耻 文明 狠毒 恶毒 残酷 野蛮 凶恶 凶残 黑暗 阴险 丑恶 残忍 残暴 刻薄 刻毒 凶暴 严酷 可耻 狂妄 凶狠 狡猾 蛮横

A6 无赖 坚定 稳 婉转 坚决 强硬 虔诚

A7 和气 红润 狠心 宏观 富有 阔

A8 耐心 积极 消极 主观 主动 顽固 保守 谨慎

A9 温厚 凌厉 凛然 磅礴

A10 贤惠 花哨 起劲 荣幸 傻 伤感 体面 内疚

8.2 Chinese noun clusters (10 of 595)

N1 老板娘 漫画 徒工 喜剧 小丑 小贩 小品 寓言

N2 阿飞 盗匪 地痞 地震 赌棍 官府 坏蛋 火灾 奸商 流氓 盲流 骗子 品性 小青年 小偷 阴谋 罪犯

N3 阿姨 见证人 奶奶 年头 售票员 信徒 助手 走狗

N4 哀乐 后母 计谋 教训 良策 内行 亲人 生路 知情人

N5 哀思 春意 光明 情操 柔情 深情 诗意

N6 爱 表格 地址 古书 经书 去向 图表 图象 渊源 帐目 哲学

N7 爱好 标语 风俗 角度 年龄 深度 体型 兴趣

N8 爱情 抱负 感情 理想 青春 情感 情谊 友谊

N9 爱人 旅客 妻子 同学

N10 案犯 部下 敌军 革命家 心态 信仰

References

[1] Donghong Ji & Changning Huang, A Semantic Composition Model for Chinese Noun and Adjective, Communications of COCIPS 6(1): P25-P33, 1996

[2] Donghong Ji, Computational Research on Issues of Lexical Semantics, Post-doctoral Research Report, Tsinghua University , P14~P26, 1997

[3] Juanzi Li et al., Two-dimensional Clustering Based on Compositional Examples, Language Engineering, P164-P169, Tsinghua University Press, 1997

[4] Hang Li & Naoki Abe， Clustering Words with the MDL Principle，cmp-lg/9605014 v2， 1996

[5] Wei Xu, The Study of Syntax-Semantics Integrated Chinese Parsing, Thesis for the Degree of Master in Computer Science, Tsinghua University, 1997

[6] Wenjie Ni et al., Modern Chinese Thesaurus, China People Press, 1984

[7] Zhaoqi Bian et al., Pattern Recognition, Tsinghua University Press, 1997