Learning New Compositions from Given Ones

Ji Donghong
Dept. of Computer Science
Tsinghua University
Beijing, 100084
P. R. China
jdh@s1000e.cs.tsinghua.edu.cn

He Jun
Dept. of Computer Science
Herbin Institute of Technology
hj@pact518.hit.edu.cn

Huang Changning
Dept. of Computer Science
Tsinghua University
Beijing, 100084
P. R. China
hcn@mail.tsinghua.edu.cn

Abstract

In this paper, we study the problem of learning new compositions of words from given ones with a specific syntactic structure, e.g., A-N or V-N structures. We first cluster words according to the given compositions, then construct a cluster-based compositional frame for each word cluster, which contains both new and given compositions relevant with the words in the cluster. In contrast to other methods, we don't pre-define the number of clusters, and formalize the problem of clustering words as a non-linear optimization one, in which we specify the environments of words based on word clusters to be determined, rather than their neighboring words. To solve the problem, we make use of a kind of cooperative evolution strategy to design an evolutionary algorithm.

1 Introduction

Word compositions have long been a concern in lexicography(Benson et al. 1986; Miller et al. 1995), and now as a specific kind of lexical knowledge, it has been shown that they have an important role in many areas in natural language processing, e.g., parsing, generation, lexicon building, word sense disambiguation, and information retrieving, etc.(e.g., Abney 1989, 1990; Benson et al. 1986; Yarowsky 1995; Church and Hanks 1989; Church, Gale, Hans, and Hindle 1989). But due to the huge number of words, it is impossible to list all compositions between words by hand in dictionaries. So an urgent problem occurs: how to automatically acquire word compositions? In general, word compositions fall into two categories: free compositions and bound compositions, i.e., collocations. Free compositions refer to those in which words can be replaced by other similar ones, while in bound compositions, words cannot be replaced freely(Benson 1990). Free compositions are *predictable*, i.e., their reasonableness can be determined according to the syntactic and semantic properties of the words in them. While bound compositions are not *predictable*, i.e., their reasonableness cannot be derived from the syntactic and semantic properties of the words in them(Smadja 1993). Now with the availability of large-scale corpus, automatic acquisition of word compositions, especially word collocations from them have been extensively studied(e.g., Choueka et al. 1988; Church and Hanks 1989; Smadja 1993). The key of their methods is to make use of some statistical means, e.g., frequencies or mutual information, to quantify the *compositional strength* between words. These methods are more appropriate for retrieving bound compositions, while less appropriate for retrieving free ones. This is because in free compositions, words are related with each other in a more loose way, which may result in the invalidity of mutual information and other statistical means in distinguishing reasonable compositions from unreasonable ones. In this paper, we start from a different point to explore the problem of automatic acquisition of free compositions. Although we cannot list all free compositions, we can select some *typical* ones as those specified in some dictionaries(e.g., Benson 1986; Zhang et al. 1994). According to the properties held by free compositions, we can reasonably suppose that selected compositions can provide strong clues for others. Furthermore we suppose that words can be classified into clusters, with the members in each cluster similar in their compositional ability, which can be characterized as the set of the words able to combined with them to form meaningful phrases. Thus any given composition, although specifying the relation between two words literally, suggests the relation between two clusters. So for each word(or clus-

Ji Donghong, He Jun and Huang Changning (1997) **Learning New Compositions from Given Ones.** In T.M. Ellison (ed.) *CoNLL97: Computational Natural Language Learning,* ACL pp 25-32.

ter), there exist some word clusters, the word (or the words in the cluster) can and only can combine with the words in the clusters to form meaningful phrases. We call the set of these clusters compositional frame of the word (or the cluster). A seemingly plausible method to determine compositional frames is to make use of pre-defined semantic classes in some thesauri(e.g., Miller et al. 1993; Mei et al. 1996). The rationale behind the method is to take such an assumption that if one word can be combined with another one to form a meaningful phrase, the words similar to them in meaning can also be combined with each other. But it has been shown that the similarity between words in meaning doesn't correspond to the similarity in compositional ability(Zhu 1982). So adopting semantic classes to construct compositional frames will result in considerable redundancy. An alternative to semantic class is word cluster based on distributional environment (Brown et al., 1992), which in general refers to the surrounding words distributed around certain word (e.g., Hatzivassiloglou et al., 1993; Pereira et al., 1993), or the classes of them(Bensch et al., 1995), or more complex statistical means (Dagan et al., 1993). According to the properties of the clusters in compositional frames, the clusters should be based on the environment, which, however, is narrowed in the given compositions. Because the given compositions are listed by hand, it is impossible to make use of statistical means to form the environment, the remaining choices are surrounding words or classes of them.

Pereira et al.(1993) put forward a method to cluster nouns in V-N compositions, taking the verbs which can combine with a noun as its environment. Although its goal is to deal with the problem of *data sparseness*, it suffers from the problem itself. A strategy to alleviate the effects of the problem is to cluster nouns and verbs simultaneously. But as a result, the problem of word clustering becomes a *bootstrapping* one, or a *non-linear* one: the environment is also to be determined. Bensch et al. (1995) proposed a definite method to deal with the generalized version of the non-linear problem, but it suffers from the problem of *local optimization*.

In this paper, we focus on A-N compositions in Chinese, and explore the problem of learning new compositions from given ones. In order to copy with the problem of sparseness, we take adjective clusters as nouns' environment, and take noun clusters as adjectives' environment. In order to avoid local optimal solutions, we propose a *cooperative evolutionary strategy*. The method uses no specific knowledge of A-N structure, and can be applied to other structures.

The remainder of the paper is organized as follows: in section 2, we give a formal description of the problem. In section 3, we discuss a kind of cooperative evolution strategy to deal with the problem. In section 4, we explore the problem of parameter estimation. In section 5, we present our experiments and the results as well as their evaluation. In section 6, we give some conclusions and discuss future work.

2 Problem Setting

Given an adjective set and a noun set, suppose for each noun, some adjectives are listed as its *compositional instances*. Our goal is to learn new reasonable compositions from the instances. To do so, we cluster nouns and adjectives simultaneously and build a compositional frame for each noun.

Suppose A is the set of adjectives, N is the set of nouns, for any $a \in A$, let $f(a) \subset N$ be the instance set of a, i.e., the set of nouns in N which can be combined with a, and for any $n \in N$, let $g(n) \subset A$ be the instance set of n, i.e., the set of adjectives in A which can be combined with n. We first give some formal definitions in the following:

Definition 1 *partition*

Suppose U is a non-empty finite set, we call $< U_1, U_2, ..., U_k >$ a *partition* of U, if:

i) for any U_i, and $U_j, i \neq j, U_i \cap U_j = \phi$

ii) $U = \cup_{1 \leq l \leq k} U_l$

We call U_i a *cluster* of U.

Suppose $\overline{U} =< A_1, A_2, ..., A_p >$ is a partition of A, $\overline{V} =< N_1, N_2, ..., N_q >$ is a partition of N, f and g are defined as above, for any N_i, let $g(N_i) = \{A_j : \exists n \in N_i, A_j \cap g(n) \neq \phi\}$, and for any n, let $\delta_{<\overline{U},\overline{V}>}(n) =| \{a : \exists A_l, A_l \in g(N_k), a \in A_j l\} - g(n) |$, where $n \in N_k$. Intuitively, $\delta_{<\overline{U},\overline{V}>}(n)$ is the number of the new instances relevant with n.

We define the *general learning amount* as the following:

Definition 2 *learning amount*

$$\delta_{<\overline{U},\overline{V}>} = \sum_{n \in N} \delta_{<\overline{U},\overline{V}>}(n)$$

Based on the partitions of both nouns and adjectives, we can define the distance between nouns and that between adjectives.

Definition 3 *distance between words*

for any $a \in A$, let $f_{\overline{V}}(a) = \{N_i : 1 \leq i \leq q, N_i \cap f(a) \neq \phi\}$, for any $n \in N$, let $g_{\overline{U}} = \{A_i : 1 \leq i \leq p, A_i \cap g(n) \neq \phi\}$, for any two nouns n_1 and n_2, any two adjectives a_1 and a_2, we define the distances between them respectively as the following:

i)

$$dis_{\overline{U}}(n_1, n_2) = 1 - \frac{| g_{\overline{U}}(n_1) \cap g_{\overline{U}}(n_2) |}{| g_{\overline{U}}(n_1) \cup g_{\overline{U}}(n_2) |}$$

ii)

$$dis_{\overline{V}}(a_1, a_2) = 1 - \frac{| f_{\overline{V}}(a_1) \cap f_{\overline{V}}(a_2) |}{| f_{\overline{V}}(a_1) \cup f_{\overline{V}}(a_2) |}$$

According to the distances between words, we can define the *distances between word sets*.

Definition 4 *distance between word sets*

Given any two adjective sets $X_1, X_2 \subset A$, any two noun sets $Y_1, Y_2 \subset N$, their *distances* are:

i)

$$dis_{\overline{V}}(X_1, X_2) = \max_{a_1 \in X_1, a_2 \in X_2} \{dis_{\overline{V}}(a_1, a_2)\}$$

ii)

$$dis_{\overline{U}}(Y_1, Y_2) = \max_{n_1 \in Y_1, n_2 \in Y_2} \{dis_{\overline{U}}(n_1, n_2)\}$$

Intuitively, the distance between word sets refer to the biggest distance between words respectively in the two sets.

We formalize the problem of clustering nouns and adjectives simultaneously as an optimization problem with some constraints.

(1)To determine a partition $\overline{U} =< A_1, A_2, ..., A_p >$ of A, and a partition $\overline{V} =< N_1, N_2, ..., N_q >$ of N, where $p, q > 0$, which satisfies i) and ii), and minimize $\delta_{<\overline{U},\overline{V}>}$.

i) for any $a_1, a_2 \in A_i, 1 \leq i \leq p$, $dis_{\overline{V}}(a_1, a_2) < t_1$; for A_i and A_j, $1 \leq i \neq j \leq p$, $dis_{\overline{V}}(A_i, A_j) \geq t_1$;

ii) for any $n_1, n_2 \in N_i, 1 \leq i \leq q$, $dis_{\overline{U}}(n_1, n_2) < t_2$; for N_i and N_j, $1 \leq i \neq j \leq p$, $dis_{\overline{U}}(N_i, N_j) \geq t_2$;

Intuitively, the conditions i) and ii) make the distances between words within clusters smaller, and those between different clusters bigger, and to minimize $\delta_{<\overline{U},\overline{V}>}$ means to minimize the distances between the words within clusters.

In fact, $(\overline{U}, \overline{V})$ can be seen as an *abstraction model* over given compositions, and t_1, t_2 can be seen as its *abstraction degree*. Consider the two special case: one is $t_1 = t_2 = 0$, i.e., the abstract degree is the lowest, when the result is that one noun forms a cluster and on adjective forms a cluster, which means that no new compositions are learned. The other is $t_1 = t_2 = 1$, the abstract degree is the highest, when a possible result is that all nouns form a cluster and all adjectives form a cluster, which means that all possible compositions, reasonable or unreasonable, are learned. So we need estimate appropriate values for the two parameters, in order to make an appropriate abstraction over given compositions, i.e., make the compositional frames contain as many reasonable compositions as possible, and as few unreasonable ones as possible.

3 Cooperative Evolution

Since the beginning of evolutionary algorithms, they have been applied in many areas in AI(Davis et al., 1991; Holland 1994). Recently, as a new and powerful learning strategy, cooperative evolution has gained much attention in solving complex non-linear problem. In this section, we discuss how to deal with the problem (1) based on the strategy.

According to the interaction between adjective clusters and noun clusters, we adopt such a cooperative strategy: after establishing the preliminary solutions, for any preliminary solution, we optimize N's partition based on A's partition, then we optimize A's partition based on N's partition, and so on, until the given conditions are satisfied.

3.1 Preliminary Solutions

When determining the preliminary population, we also cluster nouns and adjectives respectively. However, we see the environment of a noun as the set of all adjectives which occur with it in given compositions, and that of an adjective as the set of all the nouns which occur with it in given compositions. Compared with (1), the problem is a linear clustering one.

Suppose $a_1, a_2 \in A$, f is defined as above, we define the *linear distance* between them as (2):

(2) $dis(a_1, a_2) = 1 - \frac{|f(a_1) \cap f(a_2)|}{|f(a_1) \cup f(a_2)|}$

Similarly, we can define the *linear distance* between nouns $dis(n_1, n_2)$ based on g. In contrast, we call the distances in definition 3 *non-linear distances*.

According to the linear distances between adjectives, we can determine a *preliminary partition* of N: randomly select an adjective and put it into an empty set X, then scan the other adjectives in A, for any adjective in $A - X$, if its distances from the adjectives in X are all smaller than t_1, then put it into X, finally X forms a preliminary cluster. Similarly, we can build another preliminary cluster in $(A - X)$. So on, we can get a set of preliminary clusters, which is just a partition of A. According to the different order in which we scan the adjectives, we can get different preliminary partitions of A. Similarly, we can determine the preliminary partitions of N based on the linear distances between nouns. A partition of A and a partition of N forms a preliminary solution of (1), and all possible preliminary solutions forms the population of preliminary solutions, which we also call the population of *0th generation* solutions.

3.2 Evolution Operation

In general, evolution operation consists of recombination, mutation and selection. Recombination makes two solutions in a generation combine with each other to form a solution belonging to next generation. Suppose $< U_1^{(i)}, V_1^{(i)} >$ and $< U_2^{(i)}, V_2^{(i)} >$ are two *ith* generation solutions, where $U_1^{(l)}$ and $U_2^{(i)}$ are two partitions of A, $V_1^{(i)}$ and $V_2^{(i)}$ are two partitions of N, then $< U_1^{(i)}, V_2^{(i)} >$ and $< U_2^{(i)}, V_1^{(i)} >$ forms two possible *(i+1)th* generation solutions.

Mutation makes a solution in a generation improve its fitness, and evolve into a new one belonging to next generation. Suppose $< U^{(i)}, U^{(i)} >$ is a ith generation solution, where $U^{(i)} =< A_1, A_2, ..., A_p >$, $V^{(i)} =< N_1, N_2, ..., N_q >$ are partitions of A and N respectively, the mutation is aimed at optimizing $V^{(l)}$ into $V^{(l+1)}$ based on $U^{(l)}$, and makes $V^{(l+1)}$ satisfy the condition ii) in (1), or optimizing $U^{(l)}$ into $U^{(l+1)}$ based on $V^{(l)}$, and makes $U^{(l+1)}$ satisfy the condition i) in (1), then moving words across clusters to minimize $\delta_{<\overline{U},\overline{V}>}$.

We design three steps for mutation operation: *splitting, merging and moving*, the former two are intended for the partitions to satisfy the conditions in (1), and the third intended to minimize $\delta_{<\overline{U},\overline{V}>}$. In the following, we take the evolution of $V^{(l+1)}$ as an example to demonstrate the three steps.

Splitting Procedure. For any N_k, $1 \leq k \leq$, if there exist $n_1, n_2 \in N_k$, such that $dis_{U^{(l+1)}}(n_1, n_2) \geq t_2$, then splitting N_k into two subsets X and Y. The procedure is given as the following:

i) Put n_1 into X, n_2 into Y,

ii) Select the noun in $(N_k - (X \cup Y))$ whose distance from n_1 is the smallest, and put it into X,

iii) Select the noun in $(N_k - (X \cup Y))$ whose distance from n_2 is the smallest, and put it into Y,

iv) Repeat ii) and iii), until $X \cup Y = N_k$.

For X (or Y), if there exist $n_1, n_2 \in X$ (or Y), $dis_{U^{(i)}} \geq t_2$, then we can make use of the above procedure to split it into more smaller sets. Obviously, we can split any N_k in $V^{(l)}$ into several subsets which satisfy the condition ii) in (1) by repeating the procedure.

Merging procedure. If there exist N_j and N_k, where $1 \leq j, k \leq q$, such that $dis_{U^{(i)}}(N_l, N_k) < t_2$, then merging them into a new cluster.

It is easy to prove that $U^{(l)}$ and $V^{(l)}$ will meet the condition i) and ii) in (1) respectively, after splitting and merging procedure.

Moving procedure. We call moving n from N_j to N_k a word move, where $1 \leq j \neq k \leq q$, denoted as (n, N_j, N_k), if the condition (ii) remains satisfied. The procedure is as the following:

i) Select a word move (n, N_j, N_k) which minimizes $\delta_{<\overline{U},\overline{V}>}$,

ii) Move n from N_j to N_k,

iii) Repeat i) and ii) until there are no word moves which reduce $\delta_{<\overline{U},\overline{V}>}$.

After the three steps, $U^{(i)}$ and $V^{(i)}$ evolve into $U^{(i+1)}$ and $V^{(i+1)}$ respectively.

Selection operation selects the solutions among those in the population of certain generation according to their fitness. We define the fitness of a solution as its learning amount.

We use J_i to denote the set of ith generation solutions, $H(i, i+1)$, as in (3), specifies the similarity between ith generation solutions and $(i+1)th$ generation solutions.

(3)

$$H(i, i+1) = \frac{\min\{\delta_{(U^{(i+1)}, V^{(i+1)})} : (U^{(i+1)}, V^{(i+1)}) \in J_{i+1}\}}{\min\{\delta_{(U^{(i)}, V^{(i)})} : (U^{(i)}, V^{(i)}) \in J_i\}}$$

Let t_3 be a threshold for $H(i, i+1)$, the following is the general evolutionary algorithm:

Procedure Clustering(A, N, f, g);
begin
i) Build preliminary solution population I_0,
ii) Determine 0th generation solution set J_0 according to their fitness,
iii) Determine I_{i+1} based on J_i:
a) Recombination: if $(U_1^{(i)}, V_1^{(i)})$, $(U_2^{(i)}, V_2^{(i)}) \in J_i$, then $(U_1^{(i)}, V_2^{(i)})$, $(U^{(i)}, V_2^{(i)}) \in I_{i+1}$,
b) Mutation: if $(U^{(i)}, V^{(i)}) \in J_i$, then $(U^{(i)}, V^{(i+1)})$, $(U^{(i+1)}, V^{(i)}) \in I_{i+1}$,
iv) Determine J_{i+1} from I_{i+1} according to their fitness,
v) If $H(i, i+1) > t_3$, then exit, otherwise goto iii),
end

After determining the clusters of adjectives and nouns, we can construct the compositional frame for each noun cluster or each noun. In fact, for each noun cluster $N_i, g(N_i) = \{A_j : \exists n \in N_i, A_j \cap g(n) \neq \phi\}$ is just its compositional frame, and for any noun in N_i, $g(Ni)$ is also its compositional frame. Similarly, for each adjective (or adjective cluster), we can also determine its compositional frame.

4 Parameter Estimation

The parameters t_1 and t_2 in (1) are the thresholds for the distances between the clusters of A and N respectively. If they are too big, the established frame will contain more unreasonable compositions, on the other hand, if they are too small, many reasonable compositions may not be included in the frame. Thus, we should determine appropriate values for t_1 and t_2, which makes the fame contain as many reasonable compositions as possible, meanwhile as few unreasonable ones as possible.

Suppose F_i is the compositional frame of N_i, let $F =< F_1, F_2, ..., F_q >$, for any F_i, let $A_{F_i} = \{a : \exists X \in F_i, a \in X\}$. Intuitively, A_{F_i} is the set of the adjectives *learned* as the compositional instances of the noun in N_i. For any $n \in N_i$, we use A_n to denote the set of all the adjectives which in fact can modify n to form a meaningful phrase, we now define deficiency rate and redundancy rate of F. For convenience, we use δ_F to represent $\delta(\overline{U}, \overline{V})$.

Definition 5 *Deficiency rate* α_F

$$\alpha_F = \frac{\sum_{1 \leq i \leq q} \sum_{n \in N_i} \mid A_n - A_{F_i} \mid}{\sum_{n \in N} \mid A_n \mid}$$

Intuitively, α_F refers to the ratio between the reasonable compositions which are not learned and all the reasonable ones.

Definition 6 *Redundancy rate* β_F

$$\beta_F = \frac{\sum_{1 \leq i \leq q} \sum_{n \in N_i} \mid A_{F_i} - A_n \mid}{\delta_F}$$

Intuitively, β_F refers to the ratio between unreasonable compositions which are learned and all the learned ones.

So the problem of estimating t_1 and t_2 can be formalized as (5):

(5) to find t_1 and t_2, which makes $\alpha_F = 0$, and $\beta_F = 0$.

But, (5) may exists no solutions, because its constraints are two strong, on one hand, the sparseness of instances may cause α_F not to get 0 value, even if t_1 and t_2 close to 1, on the other hand, the difference between words may cause β_F not to get 0 value, even if t_1 and t_2 close to 0. So we need to weaken (5).

In fact, both α_F and β_F can be seen as the functions of t_1 and t_2, denoted as $\alpha_F(t_1, t_2)$ and $\beta_F(t_1, t_2)$ respectively. Given some values for t_1 and t_2, we can compute α_F and β_F. Although there may exist no values (t'_1, t'_2) for (t_1, t_2), such that $\alpha_F(t'_1, t'_2) = \beta_F(t'_1, t'_2) = 0$, but with t_1 and t_2 increasing, α_F tends to decrease, while β_F tends to increase. So we can weaken (5) as (6).

(6) to find t_1 and t_2, which maximizes (7).

(7)

$$\frac{\sum_{(t_1,t_2) \in \Gamma_1(t'_1,t'_2)} \alpha_F(t_1, t_2)}{\mid \Gamma_1(t'_1, t'_2) \mid} - \frac{\sum_{(t_1,t_2) \in \Gamma_2(t'_1,t'_2)} \alpha_F(t_1, t_2))}{\mid \Gamma_2(t'_1, T'_2) \mid}$$

where $\Gamma_1(t'_1, t'_2) = \{(t_1, t_2) : 0 \leq t1 \leq t'_1, 0 \leq t_2 \leq t'_2\}$, $\Gamma_2(t'_1, t'_2) = \{(t_1, t_2) : t'_1 < t1 \leq 1, t'_2 < t_2 \leq 1\}$

Intuitively, if we see the area ([0, 1]; [0, 1]) as a sample space for t_1 and t_2, $\Gamma_1(t'_1, t'_2)$ and $\Gamma_2(t'_1, t'_2)$ are its sub-areas. So the former part of (7) is the mean deficiency rate of the points in $\Gamma_1(t'_1, t'_2)$, and the latter part of (7) is the *mean deficiency rate* of the points in $\Gamma_2(t'_1, t'_2)$. To maximize (7) means to maximize its former part, while to minimize its latter part. So our weakening (5) into (6) lies in finding a point (t'_1, t'_2), such that the mean deficiency rate of the sample points in $\Gamma_2(t'_1, t'_2)$ tends to be very low, rather than finding a point (t'_1, t'_2), such that its deficiency rate is 0.

5 Experiment Results and Evaluation

We randomly select 30 nouns and 43 adjectives, and retrieve 164 compositions(see Appendix I) between them from *Xiandai Hanyu Cihai* (Zhang et al. 1994), a word composition dictionary of Chinese. After checking by hand, we get 342 reasonable compositions (see Appendix I), among which 177 ones are neglected in the dictionary. So the *sufficiency rate* (denoted as γ) of these given compositions is 47.9%.

We select 0.95 as the value of t_3, and let $t_1 = 0.0, 0.1, 0.2, ..., 1.0$, $t_2 = 0.0, 0.1, 0.2, ..., 1.0$ respectively, we get 121 groups of values for α_F and β_F. Fig.1 and Fig.2 demonstrate the distribution of α_F and β_F respectively.

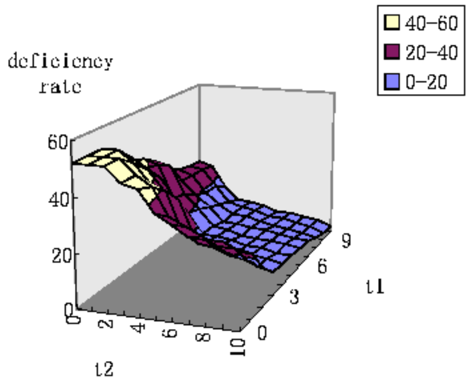

Figure 1: The distribution of α_F

For any given t_1, and t_2,we found (7) get its biggest value when $t_1 = 0.4$ and $t_2 = 0.4$, so we se-

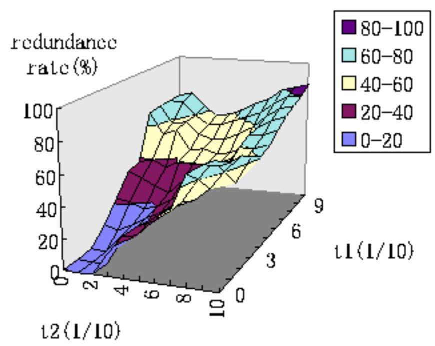

Figure 2: The distribution of β_F

$\gamma(\%)$	t_1	t_2	$\alpha_F(\%)$	$\beta_F(\%)$
32.5	0.5	0.6	13.2	34.5
47.9	0.4	0.4	15.4	26.4
58.2	0.4	0.4	10.3	15.4
72.5	0.3	0.3	9.5	7.6

Table 1: The relation between $\gamma, t_1, t_2, \alpha_F$ and β_F.

$\gamma(\%)$	$\overline{\alpha}_F(\%)$	$e_1(\%)$	$\overline{\beta}_F(\%)$	$e_2(\%)$
58.2	11.2	8.3	17.5	10.8
72.5	7.4	4.1	8.7	5.4

Table 2: The relation between γ, mean α_F and mean β_F, e_1 and e_2 is the mean error.

lect 0.4 as the appropriate value for both t_1 and t_2. The result is listed in Appendix II. From Fig.1 and Fig.2, we can see that when $t_1 = 0.4$ and $t_2 = 0.4$, both α_F and β_F get smaller values. With the two parameters increasing, α_F decreases slowly, while β_F increases severely, which demonstrates the fact that the learning of new compositions from the given ones has reached the *limit* at the point: the other reasonable compositions will be learned at a cost of severely raising the redundancy rate.

From Fig.1, we can see that α_F generally increases as t_1 and t_2 increase, this is because that to increase the thresholds of the distances between clusters means to raise the abstract degree of the model, then more reasonable compositions will be learned. On the other hand, we can see from Fig.2 that when $t_1 \geq 0.4, t_2 \geq 0.4$, β_F roughly increases as t_1 and t_2 increase, but when $t_1 < 0.4$, or $t_2 < 0.4$, β_F changes in a more confused manner. This is because that when $t_1 < 0.4$, or $t_2 < 0.4$, it may be the case that much more reasonable compositions and much less unreasonable ones are learned, with t_1 and t_2 increasing, which may result in β_F's reduction, otherwise β_F will increase, but when $t_1 \geq 0.4$, $t_2 \geq 0.4$, most reasonable compositions have been learned, so it tend to be the case that more unreasonable compositions will be learned as t_1 and t_2 increase, thus β_F increases in a rough way.

To explore the relation between γ, α_F and β_F, we reduce or add the given compositions, then estimate t_1 and t_2, and compute α_F and β_F. Their correspondence is listed in Table 1.

From Table 1, we can see that as γ increases, the estimated values for t_1 and t_2 will decrease, and β_F will also decrease. This demonstrates that if given less compositions, we should select bigger values for the two parameters in order to learn as many reasonable compositions as possible, however, which will lead to non-expectable increase in β_F. If given more compositions, we only need to select smaller values for the two parameters to learn as many reasonable compositions as possible.

We select other 10 groups of adjectives and nouns, each group contains 20 adjectives and 20 nouns. Among the 10 groups, 5 groups hold a sufficiency rate about 58.2%, the other 5 groups a sufficiency rate about 72.5%. We let $t_1 = 0.4$ and $t_2 = 0.4$ for the former 5 groups, and let $t_1 = 0.3$ and $t_2 = 0.3$ for the latter 5 groups respectively to further consider the relation between γ, α_F and β_F, with the values for the two parameters fixed.

Table 2 demonstrates that for any given compositions with fixed sufficiency rate, there exist close values for the parameters, which make α_F and β_F maintain lower values, and if given enough compositions, the mean errors of α_F and β_F will be lower. So if given a large number of adjectives and nouns to be clustered, we can extract a small sample to estimate the appropriate values for the two parameters, and then apply them into the original tasks.

6 Conclusions and Future work

In this paper, we study the problem of learning new word compositions from given ones by establishing compositional frames between words. Although we focus on A-N structure of Chinese, the method uses no structure-specific or language-specific knowledge, and can be applied to other syntactic structures, and other languages.

There are three points key to our method. First, we formalize the problem of clustering adjectives and nouns based on their given compositions as a non-linear optimization one, in which we take noun clusters as the environment of adjectives, and adjective clusters as the environment of nouns. Second, we design an evolutionary algorithm to determine its optimal solutions. Finally, we don't pre-define the number of the clusters, instead it is automatically determined by the algorithm.

Although the effects of the sparseness problem can be alleviated compared with that in traditional methods, it is still the main problem to influence the learning results. If given enough and typical compositions, the result is very promising. So important future work is to get as many typical compositions as possible from dictionaries and corpus as the foundation of our algorithms.

At present, we focus on the problem of learning compositional frames from the given compositions with a single syntactic structure. In future, we may take into consideration several structures to cluster words, and use the clusters to construct more complex frames. For example, we may consider both A-N and V-N structures in the meantime, and build the frames for them simultaneously.

Now we make use of sample points to estimate appropriate values for the parameters, which seems that we cannot determine very accurate values due to the computational costs with sample points increasing. Future work includes how to model the sample points and their values using a continuous function, and estimate the parameters based on the function.

References

Abney, S. 1989. Parsing by Chunks. In C. Tenny ed. *The MIT Parsing Volume*, MIT Press.

Abney, S. 1990. Rapid Incremental Parsing with Repair. in *Proceedings of Waterloo Conference on Electronic Text Research.*

Bensch, P.A. and W. J. Savitch. 1995. An Occurrence-Based Model of Word Categorization, *Annals of Mathematics and Artificial Intelligence,* 14:1-16.

Benson, M., Benson, E., and Ilson, R. 1986. *The lexicographic Description of English.* John Benjamins.

Benson, M. 1986. *The BBI Combinatory Dictionary of English: A Guide to Word Combinations.* John Benjamins.

Benson. M. 1990. Collocations and General - Purpose Dictionaries. *International Journal of Lexicography,* 3(1): 23-35.

Davis, L. et al. 1991. *Handbook of Genetic Algorithms.* New York: Van Nostrand, Reinhold.

Choueka, Y., T. Klein, and E. Neuwitz. 1983. Automatic Retrieval of Frequent Idiomatic and Collocational Expressions in a Large Corpus. *Journal of Literary and Linguistic Computing,* 4: 34-38.

Church, K. and P. Hanks. 1989. Word Association Norms, Mutual Information, and Lexicography, in *Proceedings of 27th Annual Meeting of the Association for Computational Linguistics,* 76-83.

Church, K., W. Gale, P. Hanks, and D. Hindle. 1989. Parsing, Word Associations and Typical Predicate-Argument relations, in *Proceedings of the International Workshop on Parsing Technologies*, Carnegie Mellon University, Pittsburgh, PA. 103-112.

Holland, J.H. 1992. *Adaption in Natural and Artificial Systems*, 2nd edition, Cambridge, Massachusetts, MIT Press.

Hatzivassiloglou, V. and K.R.Mckeown. Towards the Automatic Identification of Adjectival Scales: Clustering of adjectives According to Meaning. In *Proceedings of Annual Meeting of 31st ACL,* Columbus, Ohio, USA.

Lin, X.G. et al. 1994. *Xiandai Hanyu Cihai.* Renmin Zhongguo Press(in Chinese).

Mei, J.J. et al. 1983. *TongYiCi CiLin (A Chinese Thesaurus).* Shanghai Cishu press, Shanghai.

Miller, G.A., R. Backwith, C. Fellbaum, D. Gross, K.J. Miller. 1993 Introduction to WordNet: An On-line Lexical Database, International *Journal of Lexicography,* (Second Edition.

Pereira, F., N. Tishby, and L. Lillian. 1993. Distributional Clustering of English Words, In *Proceedings of Annual Meeting of 31st ACL,* Columbus, Ohio, USA, 1995.

Smadia, F. 1993. Retrieving Collocations from Text: Xtract, *Computational Linguistics,* 19(1).

Yarowsky, D. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33th Annual Meeting of the Association for Computational Linguistics,* Cambridge, Massachusetts.

Zhu, D.X. 1982. *Lectures in Grammar.* Shanghai Education Press(in Chinese).

Appendix I

In this appendix, we listed the 30 nouns, and for any one of the nouns, we also list the adjectives which can combined with it to form a meaningful phrase5.

1 友谊：珍贵真挚诚挚宝贵可贵 // 美好美妙迷人

2 友情：真挚宝贵珍贵可贵诚挚 // 美好美妙迷人

3 田野：广阔宽广美丽迷人可爱 // 宽阔美妙

4 感情：真挚美好健康可贵美丽诚挚悲伤 // 宝贵珍贵美妙迷人紧张愉快痛苦沮丧苦闷郁闷烦闷疲惫

5 原野：美丽迷人广阔明媚宽阔宽广 // 美妙可爱

6 爱情：美好珍贵真挚美妙健康迷人宝贵可贵 // 美丽诚挚欢乐愉快悲伤沮丧痛苦紧张苦闷郁闷烦闷疲惫

7 技术：熟练娴熟 // 珍贵宝贵可贵

8 心情：愉快沮丧欢乐悲痛烦闷苦闷不安懊悔懊恼紧张乐观悲伤 // 疲惫疲劳痛苦悲痛郁闷烦闷

9 神色：紧张不安痛苦懊恼疲惫懊悔 // 美丽美妙美好迷人可爱懊悔疲劳疲乏烦闷苦闷郁闷欢乐愉快乐观悲伤沮丧悲痛健康

10 情绪：健康愉快沮丧烦闷不安乐观疲劳懊悔懊恼紧张 // 悲痛痛苦苦闷欢乐悲伤疲倦疲惫疲乏郁闷

11 阳光：明丽明媚迷人 // 美丽美妙美好可爱

12 春光：明媚美好迷人明丽 // 美丽美妙可爱

13 春色：美妙美好迷人 // 美丽可爱明媚明丽

14 情谊：真挚宝贵美好珍贵可贵诚挚 // 美妙迷人

15 生活：愉快美好痛苦艰难困难美丽美妙紧张欢乐 // 迷人可爱艰巨不安烦闷苦闷郁闷乐观悲伤悲痛沮丧健康

16 技巧：娴熟熟练 // 宝贵珍贵可贵迷人

17 青春：美好美丽迷人欢乐可爱可贵愉快美妙宝贵珍贵

18 年华：美好宝贵美丽可贵 // 珍贵美妙迷人可爱愉快欢乐

19 身体：健康疲惫疲乏疲劳

20 身子：健康疲惫疲劳疲乏

21 信心：坚定坚强 // 宝贵珍贵可贵顽强

22 信念：坚定坚强美好真挚 // 顽强美妙诚挚宝贵可贵珍贵

23 心境：愉快懊悔不安宽阔苦闷懊恼 // 紧张宽广广阔烦闷苦闷欢乐乐观悲伤沮丧美好美妙迷人悲痛痛苦

24 心灵：美丽美好宽广悲痛痛苦广阔苦闷 // 烦闷健康郁闷可爱迷人悲伤不安

25 心胸：宽广宽阔广阔 // 烦闷苦闷欢乐乐观悲伤沮丧愉快

26 任务：艰巨困难 // 艰难紧张

27 毅力：坚强顽强 // 坚定宝贵珍贵可贵

28 意志：坚强坚定顽强 // 宝贵珍贵可贵

29 性格：美好坚强坚定可爱顽强乐观 // 迷人美妙烦闷苦闷欢乐悲伤沮丧宝贵珍贵可贵郁闷

30 神情：懊悔懊恼疲乏疲惫疲劳紧张不安苦闷悲伤沮丧 // 烦闷郁闷痛苦悲痛愉快欢乐乐观悲伤健康美妙美好迷人可爱美丽

Appendix II

1) lists noun clusters and their compositional frames, 2) lists adjective clusters.

1). Noun Clusters and Their Compositional Frames:

N1 友谊友情情谊： A1 A10
N2 感情爱情： A1 A2 A10 A11 A13
N3 田野原野： A2 A3 A9
N4 技术技巧： A12
N5 神色神情情绪： A4 A5 A8 A11
N6 阳光春光春色： A2 A3
N7 生活： A2 A7 A8 A11 A13 A14
N8 青春年华： A1 A2 A14
N9 身体身子： A4 A13
N10 信心信念性格： A2 A6 A10 A14
N11 心境心灵心胸： A2A8 A9 A10A11
N12 毅力意志： A6
N13 任务： A7
N14 性格： A2 A6 A14

2). Adjective Clusters:

A1 宝贵珍贵可贵
A2 美丽美妙美好迷人可爱
A3 明媚明丽
A4 疲惫疲乏疲劳
A5 懊悔懊恼沮丧
A6 坚定坚强顽强
A7 艰巨艰难困难
A8 紧张不安
A9 广阔宽广宽阔
A10 真挚诚挚
A11 烦闷郁闷苦闷 悲伤 痛苦 悲痛
A12 娴熟 熟练
A13 健康
A14 欢乐 愉快 乐观

1 The compositional instances of the adjectives can be inferred from those of the nouns.

2 Sufficiency rate refers to the ratio between given reasonable compositions and all reasonable ones.

3 On some points, it may be not the case.

4 For a variable X, suppose its value are X_1, X_2, $\cdots$, X_n, its mean error refers to .

5 The adjectives before "//" are those retrieved from the word composition dictionary, and those after "//" are those added by hand.