Learning New Compositions from Given Ones

Ji Donghong
Dept. of Computer Science
Tsinghua University
Beijing, 100084
P. R. China
jdh@1000e.cs.tsinghua.edu.cn

He Jun
Dept. of Computer Science
Herbin Institute of Technology
Beijing, 100084
P. R. China
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-
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 classes 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 \( f(a) \subseteq N \) 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 \( g(n) \subseteq A \) 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 \( U_j, i \neq j, U_i \cap U_j = \emptyset \)

ii) \( U = U_1 \cup U_2 \cup ... \cup U_k \)

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 \( g(N_i) = \{ A_j : \exists n \in N_i, A_j \cap g(n) \neq \emptyset \} \), and for any \( n \), let \( \delta_{<\overline{U}, \overline{V}>}(n) = |\{ a : \exists A_i, A_j \in (N_k), a \in A_j \} - g(n) | \), 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 \( f(a) = \{ N_i : 1 \leq i \leq q, N_i \cap f(a) \neq \emptyset \} \), for any \( n \in N \), let \( g(n) = \{ A_i : 1 \leq i \leq p, A_i \cap g(n) \neq \emptyset \} \), 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:
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) $\text{dis}_V(X_1, X_2) = \max_{a_1 \in X_1, a_2 \in X_2} \{\text{dis}_V(a_1, a_2)\}$

ii) $\text{dis}_G(Y_1, Y_2) = \max_{n_1 \in Y_1, n_2 \in Y_2} \{\text{dis}_G(n_1, n_2)\}$

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

i) for any $a_1, a_2 \in A_1, 1 \leq i \leq p, \text{dis}_V(a_1, a_2) < t_1$;
   for $A_i$ and $A_j$, $1 \leq i \neq j \leq p, \text{dis}_G(A_i, A_j) \geq t_1$;

ii) for any $n_1, n_2 \in N_1, 1 \leq i \leq q, \text{dis}_G(n_1, n_2) < t_2$;
   for $N_i$ and $N_j$, $1 \leq i \neq j \leq p, \text{dis}_G(N_i, N_j) \geq t_2$;

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

In fact, $(U, V)$ 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 cases: 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):

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

Similarly, we can define the linear distance between nouns $\text{dis}(n_1, n_2)$ 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 \( i \)th 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 \( \langle U_1(i), V_1(i) \rangle \) and \( \langle U_2(i), V_2(i) \rangle \) are two \( i \)th generation solutions, where \( U_1(i) \) and \( U_2(i) \) are two partitions of \( A \), \( V_1(i) \) and \( V_2(i) \) are two partitions of \( N \), then \( \langle U_1(i), V_2(i) \rangle \) and \( \langle U_2(i), V_1(i) \rangle \) 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 \( \langle U(i), U(i) \rangle \) is a \( i \)th generation solution, where \( U(i) = \langle A_1, A_2, \ldots, A_p \rangle \), \( V(i) = \langle N_1, N_2, \ldots, N_q \rangle \) are partitions of \( A \) and \( N \) respectively, the mutation is aimed at optimizing \( V(i) \) into \( V(i+1) \) based on \( U(i) \), and makes \( V(i+1) \) satisfy the condition ii) in (1), or optimizing \( U(i) \) into \( U(i+1) \) based on \( V(i) \), and makes \( U(i+1) \) satisfy the condition i) in (1), then moving words across clusters to minimize \( \delta_{U,V} \).

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 \( \delta_{U,V} \). In the following, we take the evolution of \( V(i+1) \) as an example to demonstrate the three steps.

**Splitting Procedure.** For any \( N_k \), \( 1 \leq k \leq q \), if there exist \( n_1, n_2 \in N_k \), such that \( \text{dis}(U(i),V(i)) \geq t_2 \), 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 \((X \cup Y)\), if there exist \( n_1, n_2 \in X \) or \( Y \), \( \text{dis}(U(i),V(i)) \geq t_2 \), then we can use the manipulation of the above procedure to split it into more smaller sets. Obviously, we can split any \( N_k \) in \( V(i) \) into several subsets which satisfy the condition ii) in (1) by repeating the procedure.

**Merging procedure.** If there exist \( N_i \) and \( N_k \), where \( 1 \leq j, k \leq q \), such that \( \text{dis}(U(i),N_i, N_k) < t_2 \), then merging them into a new cluster.

It is easy to prove that \( U(i) \) and \( V(i) \) 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 \( \delta_{U,V} \).

ii) Move \( n \) from \( N_j \) to \( N_k \).

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

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 \( i \)th generation solutions, \( H(i, i+1) \), as in (3), specifies the similarity between \( i \)th generation solutions and \((i + 1)\)th generation solutions.

\[
H(i, i+1) = \frac{\min \{ \delta(U(i),V(i+1)) : (U(i), V(i+1)) \in J_{i+1} \}}{\min \{ \delta(U(i),V(i)) : (U(i), V(i)) \in J_i \}}
\]

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 \( 0 \)th generation solution set \( J_0 \) according to their fitness,

iii) Determine \( I_{i+1} \) based on \( J_i \):

a) Recombination: if \( \langle U_1(i),V_1(i) \rangle \), \( \langle U_2(i),V_2(i) \rangle \) \( \in J_i \), then \( \langle U(i),V(i) \rangle \), \( \langle U(i),V(i) \rangle \) \( \in I_{i+1} \).

b) Mutation: if \( \langle U(i),V(i) \rangle \) \( \in J_i \), then \( \langle U(i),V(i+1) \rangle \), \( \langle U(i),V(i+1) \rangle \) \( \in I_{i+1} \).

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 \( N_i \), \( g(N_i) \) 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 \( N \) re-
spectively. 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 frame 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 $F = \langle F_1, F_2, ..., F_q \rangle$, 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 $\delta$ to represent $\delta(U, V)$.

**Definition 5** Deficiency rate $\alpha_F$

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

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

**Definition 6** Redundancy rate $\beta_F$

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

Intuitively, $\beta_F$ 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 $\alpha_F$ 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 $\beta_F$ not to get 0 value, even if $t_1$ and $t_2$ close to 0. So we need to weaken (5).

In fact, both $\alpha_F$ and $\beta_F$ 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 $\alpha_F$ and $\beta_F$. 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, $\alpha_F$ tends to decrease, while $\beta_F$ tends to increase. So we can weaken (5) as (6).

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

$$\sum_{(t_1, t_2) \in \Gamma_1(t_1', t_2')} \alpha_F(t_1, t_2)$$

where $\Gamma_1(t_1', t_2') = \{ (t_1, t_2) : 0 \leq t_1 \leq t_1', 0 \leq t_2 \leq t_2' \}$, $\Gamma_2(t_1', t_2') = \{ (t_1, t_2) : t_1' < t_1 \leq 1, t_2' < t_2 \leq 1 \}$

Intuitively, if we see the area $[0, 1] \times [0, 1]$ 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-regions. 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 $\gamma$) 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 $\alpha_F$ and $\beta_F$. Fig.1 and Fig.2 demonstrate the distribution of $\alpha_F$ and $\beta_F$ respectively.

![Figure 1: The distribution of $\alpha_F$](image)

For any given $t_1$, and $t_2$, we found (7) gets its biggest value when $t_1 = 0.4$ and $t_2 = 0.4$, so we se-
We select 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 $\alpha_F$ and $\beta_F$ get smaller values. With the two parameters increasing, $\alpha_F$ decreases slowly, while $\beta_F$ 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 $\alpha_F$ 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 $t_1 \geq 0.4, t_2 \geq 0.4$, $\beta_F$ roughly increases as $t_1$ and $t_2$ increase, but when $t_1 < 0.4$, or $t_2 < 0.4$, $\beta_F$ 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 $\beta_F$'s reduction, otherwise $\beta_F$ 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 $\beta_F$ increases in a rough way.

To explore the relation between $\gamma$, $\alpha_F$ and $\beta_F$, we reduce or add the given compositions, then estimate $t_1$ and $t_2$, and compute $\alpha_F$ and $\beta_F$. Their correspondence is listed in Table 1.

From Table 1, we can see that as $\gamma$ increases, the estimated values for $t_1$ and $t_2$ will decrease, and $\beta_F$ 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 $\beta_F$. 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 $\gamma$, $\alpha_F$ and $\beta_F$, 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 $\alpha_F$ and $\beta_F$ maintain lower values, and if given enough compositions, the mean errors of $\alpha_F$ and $\beta_F$ 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 nonlinear 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 com-
positions, the result is very promising. So important
future work is to get as many typical compositions
as possible from dictionaries and corpus as the foun-
dation 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 com-
plex 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 in-
creasing. 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 In-
telligence, 14:1-16.

Benson, M., Benson, E., and Ilson, R. 1986. The lexi-
ographic Description of English. John Ben-
jamins.

Benson, M. 1986. The BBI Combinatory Dictionary
of English: A Guide to Word Combinations. John Ben-
jamins.

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

Davis, L. et al. 1991. Handbook of Genetic Algo-
rithms. New York: Van Nostrand, Reinhold.

Choueka, Y., T. Klein, and E. Neuwitz. 1983. Au-
tomatic 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 Lexicogra-
phy, 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 Typi-
cal Predicate-Argument relations, in Proceed-
ings of the International Workshop on Pars-
ing Technologies, Carnegie Mellon University,
Pittsburgh, PA. 103-112.

Holland, J.H. 1992. Adaptation in Natural and Arti-
ficial Systems, 2nd edition, Cambridge, Mas-
sachusetts, 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. Dis-
tributional 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 disam-
biguation rivaling supervised methods. In Pro-
cedings of the 33th Annual Meeting of the As-
sociation for Computational Linguistics, Cam-
brIDGE, 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 phrase.

1. 友谊：珍贵友谊是挚爱可贵 // 美好友谊可贵
2. 友情：真挚友谊是可贵 // 美好友谊可贵
3. 田野：广阔田野是美妙 // 美好田野可贵
4. 感情：真挚感情是美妙可贵 // 美好感情可贵
5. 原野：美丽原野是广阔 // 美好原野可贵
6. 爱情：美好爱情是美妙 // 美好爱情可贵
7. 技术：熟练技艺 // 珍贵技艺可贵
8. 心情：愉快心情是悲伤 // 美好心情可贵
9. 神色：紧张神色是可爱 // 美好神色可贵
10. 情绪：健康情绪是悲伤 // 美好情绪可贵
11. 阳光：明媚阳光 // 美好阳光可贵
12. 春光：明媚春光 // 美好春光可贵
13. 春色：美好春色 // 美好春色可贵
14. 情谊：真挚情谊是可爱 // 美好情谊可贵
15. 生活：愉快生活是悲伤 // 美好生活可贵
16. 技巧：娴熟技艺 // 宝贵技艺可贵
17. 青春：美好青春 // 美好青春可贵
18. 年华：美好年华 // 珍贵年华可贵
19. 身体：健健康康 // 珍贵身体可贵
20. 身子：健健康康 // 珍贵身体可贵
21. 信心：坚定信心 // 宝贵信心可贵
22. 信念：坚定信念 // 宝贵信念可贵
23. 心境：愉快心境 // 宝贵心境可贵
24. 灵魂：美好灵魂 // 珍贵灵魂可贵
25. 心肺：宽广胸膛 // 难过心情可贵
26. 任务：艰巨任务 // 可贵任务
27. 毅力：坚强毅力 // 宝贵毅力可贵
28. 意志：坚定意志 // 宝贵意志可贵
29. 性格：美好性格 // 宝贵性格可贵
30. 性格：坚强性格 // 宝贵性格可贵

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

1. Noun Clusters and Their Compositional Frames:

N1) 珍贵友谊
d1) A1 A10
N2) 美好爱情
d2) A1 A2 A10
N3) 珍贵技艺
d3) A2 A3 A9
N4) 真挚情谊
d4) A2 A3 A9
N5) 珍贵心境
d5) A4 A5 A8 A11
N6) 美好灵魂
d6) A2 A3 A9
N7) 悲伤心情
d7) A2 A7 A8 A11
N8) 珍贵身体
d8) A1 A2 A14
N9) 珍贵身体
d9) A4 A13
N10) 珍贵情绪
d10) A2 A6 A10
N11) 珍贵心情
d11) A2 A8 A9 A10

2. Adjective Clusters:

A1) 珍贵A1)
A2) 悲伤A2)
A3) 健康A3)
A4) 坚定A4)
A5) 珍贵A5)
A6) 宝贵A6)
A7) 珍贵A7)
A8) 健康A8)
A9) 宝贵A9)
A10) 健康A10)
A11) 美好A11)
A12) 悲伤A12)
A13) 珍贵A13)
A14) 宝贵A14)

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 X1, X2, ..., Xn, 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.

Appendix II

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

1. Noun Clusters and Their Compositional Frames:

N1) 友谊友谊
d1) A1 A10
N2) 感情爱情
d2) A1 A2 A10
N3) 真挚情谊
d3) A2 A3 A9
N4) 技术技巧
d4) A12
N5) 神色神情
d5) A4 A5 A8 A11
N6) 阳光光色
d6) A2 A3
N7) 活性生活
d7) A2 A7 A8 A11
N8) 青春年华
d8) A1 A2 A14
N9) 身体身体
d9) A4 A13
N10) 心境心情
d10) A2 A6 A10 A14
N11) 紧张心情
d11) A2 A8 A9 A10

2. Adjective Clusters:

A1) 珍贵A1)
A2) 美好A2)
A3) 明媚A3)
A4) 疲劳A4)
A5) 悲伤A5)
A6) 坚定A6)
A7) 难过A7)
A8) 紧张A8)
A9) 宽广A9)
A10) 宝贵A10)
A11) 难过A11)
A12) 紧张A12)
A13) 健康A13)
A14) 宝贵A14)