MFC: A Method of Co-referent Relation Acquisition from Large-Scale Chinese Corpora*

Guogang Tian[1,2], Cungen Cao[1], Lei Liu[1,2], and Haitao Wang[1,2]

[1] Key Laboratory of Intelligent Information Processing
Institute of Computing Technology, Chinese Academy of Sciences
[2] Graduate University of the Chinese Academy of Sciences
Beijing, China, 100080
naitgg@hotmail.com, cgcao@ict.ac.cn, liuliu_leilei@ict.ac.cn,
htwang@ict.ac.cn

Abstract. This paper proposes a multi-feature constrained method (MFC) to acquire co-referent relations from large-scale Chinese corpora. The MFC has two phases: candidate relations extraction and verification. The extraction phase uses distribution distance, pattern homogeneity and coordination distribution features of co-referent target words to extract candidate relations from Chinese corpora. In the verification phase, we define an ontology for co-referent token words, and build a relation graph for all candidate relations. Both the ontology and the graph are integrated to generate individual, joint and reinforced strategies to verify candidate relations. Comprehensive experiments have shown that the MFC is practical, and can also be extended to acquire other types of relations.

1 Introduction

In the field of knowledge acquisition from text (KAT), relation acquisition mainly focuses on hyponymy and meronymy. Hearst [1] used lexicon-syntactic patterns to acquire hyponymy from corpora. Cederberg et al. [2] applied latent semantic analysis and coordination patterns to simultaneously increase precision and recall rates of hyponymy acquisition. Girju et al. [3] defined three patterns *Y verb(has) X*, *Y's X* and *X of Y* to acquire meronymy. Berland et al. [4] used statistical method to acquire meronymic relations from large-scale corpora.

Different from the above works, Maedche et al. [5] acquired non-category relations from domain text using association rule algorithm. Lin et al. [6] computed contextual similarity to find similar words from corpora on the basis of dependency tuples.

Co-referent relations are particular relations between words, and they relate two or more words that have identical meanings and designate the same entity in the real world. Based on co-referent relations' semantics and distributions in corpora, this paper proposes a multi-feature constrained method (MFC) to acquire co-referent relations from large-scale Chinese corpora. To our knowledge, there has been little work on the

* This work is supported by the Natural Science Foundation (grant nos. 60273019, 60496326, 60573063, and 60573064), and the National 973 Programme (grants no. 2003CB317008 and G1999032701).

L. Wang et al. (Eds.): FSKD 2006, LNAI 4223, pp. 1259 – 1268, 2006.

co-referent relation acquisition so far. The acquisition is divided into two phases: candidate relations extraction and verification. The applied features include distributive features, semantic features and structural features. The distributive features refer that which position's words in a sentence can constitute a co-referent relation. The semantic features refer to the relations themselves' meaning. The structural features refer to the composition structure while all co-referent relations are linked together.

The rest of this paper is organized as follows. Section 2 and 3 introduce the extraction and verification of the co-referent relations, respectively. The experiments are presented in section 4. Section 5 concludes the paper.

2 A Hybrid Method for Extracting Co-referent Relations

In Chinese, co-referent relations may have a number of token words, e.g. 简称 (Abbreviation), 全称 (Full-Name), 又称 (Other-Name), 俗称 (Colloquialism), 学名 (Academic-Name), and 英文名称 (English-Name). They are called co-referent token words. A token word represents a type of co-referent relation.

Definition 1: Given a word set $W=\{w_1, \dots, w_n\}$, for any w_i, $w_j \in W$ ($1 \le i, j \le n$), if w_i and w_j nominate an identical entity, then w_i and w_j satisfy a co-referent relation.

Definition 2: Given a word set $W=\{w_1, \dots, w_n\}$ and a co-referent token word set $\mathcal{R}=\{\mathcal{R}_1, \dots, \mathcal{R}_m\}$, for any w_i, $w_j \in W$ ($1 \le i, j \le n$), there exists $\mathcal{R}_k \in \mathcal{R}$ ($1 \le k \le m$), if they satisfy a co-referent relation whose token is $\mathcal{R}_k$, then they are represented as $\mathcal{R}_k(w_i, w_j)$, among which $\mathcal{R}_k$ is a token word, and w_i and w_j are called target words (see Fig.2).

To extract candidate co-referent relations from the corpora, we introduce word distribution distance as a fundamental measure, and word pattern homogeneity and coordinate distribution as its supplements.

2.1 Word Distribution Distance

Word distribution distance is based on a hypothesis that co-referent relations often occur in the local sentence, centered by token words, not spread the whole sentence.

Given a sentence $S_k=w_n^-\dots w_i^-\dots w_1^-(RT)w_1^+\dots w_j^+\dots w_m^+$ in the co-referent corpora, where RT is a token word, w_i^- is the i^{th} target word in the left context of RT (from right to left), and w_j^+ is the j^{th} target word in the right context of RT (from left to right). For w_i^- and w_j^+, their relative positions to RT in S_k are denoted as $pos(w_i^-, RT)=i$ and $pos(w_j^+, RT)=j$, respectively. And their relative distances to RT in S_k are formulated in equation (1) and (2). $Len(x)$ represents the length of string x (Note: A Chinese character is 2 long.). For a word w, its distribution distance $Ddis(w, RT)$ in the corpora is defined in equation (3).

$$dis(w_i^-, RT) = \sum_{k=1}^{i} Len(w_k^-) \ . \tag{1}$$

$$dis\,(w_j{}^{+}, RT\,) = \sum_{k=1}^{j} Len\,(w_k{}^{+})\ . \tag{2}$$

$$Ddis(w,RT) = \sum_{p}\sum_{d} P(pos(w,RT) = p) \times P(dis(w,RT) = d) \times d\ . \tag{3}$$

In equation (3),

$$P(pos(w,RT\,) = p) = \frac{N(pos(w,RT\,) = p)}{\sum_{k} N(pos(w,RT\,) = k)}\ . \tag{4}$$

$$P(dis(w,RT) = d) = \frac{N(dis(w,RT) = d)}{\sum_{l} N(dis(w,RT) = l)}\ . \tag{5}$$

$P(pos(w, RT)=p)$ is the probability of that the relative position of w is p. $P(dis(w, RT)=d)$ is the probability of that the relative distance of w is d. $N(pos(w, RT)=p)$ is the occurrence number of w at the relative position p. $N(dis(w, RT)=d)$ is the occurrence number of w whose relative distance is d.

Some explanation is needed for equation (3). Even though w occurs at the same position in different sentences, the relative distance is possibly different in them. Therefore, its contribution to the distribution distance is different. Even though the relative distance of w is the same, it possibly occurs at different positions. Thus, the relative position's contribution to the distribution distance is different. Based on these considerations, we use the probabilities of the relative position and distance as the weight to reveal their contributions to the distribution distance.

In a sentence, a target word w is possibly before or after a token word RT. So w has two distribution distances: *forward distribution distance* and *backward distribution distance*. If the following three conditions are simultaneously satisfied, w_i^- and w_j^+ can form a candidate co-referent relation $\mathcal{R}_k(w_i^-, w_j^+)$:

- $Ddis(w_i^-, RT) \leq -d^l$;
- $Ddis(w_j^+, RT) \leq +d^r$;
- $Ddis(w_i^-, RT) + Ddis(w_j^+, RT) \leq d_{t.}$;

where $-d^l$ is the threshold of the backward distribution distance, $+d^r$ is the threshold of the forward distribution distance, d_t is the threshold of the sum of two distances.

2.2 Pattern Homogeneity of Target Words

Word patterns are classified into composition patterns and context patterns. In the MFC, a composition pattern refers to the POS (part of speech) pattern of a target word, and the context pattern refers to paired symbols surrounding a target word, such as quotation marks and book marks (e.g. 《 and 》 used in Chinese).The word pattern homogeneity means that two target words have the same pattern, i.e. composition or context pattern. For any two target words separated by token words in a sentence,

(1) If they have the same affix (prefix or suffix), or the same POS in the head or tail, then their composition patterns are homogeneous, and they can form a candidate co-referent relation.

For example, in the sentence 囊状淋巴管瘤好发于颈部, 又称囊状水瘤 (*Cystic lymphangioma*, *also called hygroma*, *often occurs in the neck*), 囊状淋巴管瘤(*cystic lymphangioma*) and 囊状水瘤(*hygroma*) have the same prefix (i.e. 囊状) and suffix(i.e. 瘤), so they can form a candidate relation 又称(囊状淋巴管瘤, 囊状水瘤), i.e. *Other-Name* (*cystic lymphangioma*, *hygroma*).

(2) If they have identical paired symbols, then their context patterns are homogeneous, and they can form a candidate co-referent relation.

For example, in the sentence "疯羊病"俗称"羊瘙痒病" (*Mad sheep disease has a colloquialism sheep scrapies.*), 疯羊病(*mad sheep disease*) and 羊瘙痒病(*sheep scrapies*) have the same context pattern that are both enclosed by double quotation marks. Thus, they can form a candidate relation 俗称(疯羊病, 羊瘙痒病), i.e. *colloquialism* (*mad sheep disease*, *sheep scrapies*).

2.3 Target Words Coordination

Given a word set W, for any w_i, $w_j \in W$, if they satisfy the following constraints:

- Occurring in the same sentence S_k;
- Being separated by what words, such as "、" (a coordinate sign in Chinese), "和 (*and*)" and "与 (*and*)", can express a coordinate relation;

Then words in W are considered to have coordinate relation. Based on it, we propose a coordinate distribution strategy to generate more candidate relations.

Given a word w and a coordinate set W, if there exist $w_i \in W$ such that $\mathcal{R}^c_k$ (w, w_i) holds, then for any other word $w_j \in W$, $\mathcal{R}^c_k$ (w, w_j) also holds.

For example, in the sentence 月饼，又称宫饼、小饼、月团、团圆饼等 (*Moon cake is also called palace cake*, *small cake*, *moon paste*, *and Mid-Autumn cake*), 宫饼 (*palace cake*), 小饼(*small cake*), 乐团(*moon paste*) and 团圆饼(*Mid-Autumn cake*) are coordinate in the sentence. If a candidate relation 又称(月饼, 宫饼) (i.e. *Other-Name* (*moon cake*, *palace cake*))is known in advance, the MFC can generate other three candidates, that is

又称(月饼, 小饼), i.e. *Other-Name* (*moon cake*, *small cake*)
又称(月饼, 月团), i.e. *Other-Name* (*moon cake*, *moon paste*)
又称(月饼, 团圆饼), i.e. *Other-Name* (*moon cake*, *Mid-Autumn cake*)

3 Co-referent Relations Verification

Errors may exist in the candidate relations set. The MFC must verify the candidate set to remove incorrect ones. To do this, we define a co-referent relation ontology and graph that the former reflects the semantic features, and the latter reflects the structural features.

3.1 The Ontology of Co-referent Relations

The co-referent relation ontology defines co-referent token words and their interrelations, and constraint axioms derived from them. This section mainly discusses the constraint axioms that are very important to the verification.

I. Word's Length Inequality Axioms

(A) $\forall w_i, w_j \in W, \exists \mathcal{R}^c_p \in \mathcal{R}^c, \mathcal{R}^c_p(w_i, w_j) \rightarrow Len(w_i) > Len(w_j) \vee Len(w_i) < Len(w_j)$;

For example, $Abbreviation(w_i, w_j) \rightarrow Len(w_i) > Len(w_j)$ and $Full\text{-}Name(w_i, w_j) \rightarrow Len(w_i) < Len(w_j)$.

II. Language Distribution Axioms

(B) $\forall w_i, w_j \in W, \exists \mathcal{R}^c_p \in \mathcal{R}^c, \mathcal{R}^c_p(w_i, w_j) \rightarrow ContainLang(w_i, \text{LANGTOKEN}) \vee ContainLang(w_j, \text{LANGTOKEN})$;

$ContainLang(w_i, \text{LANGTOKEN})$ means that the word w_i must contain a certain language. LANGTOKEN is a system-defined token. For example ENG refers to English, and CHN refers to Chinese. For example, $Is\text{-}Chinese\text{-}Name\ (w_i, w_j) \rightarrow ContainLang\ (w_j, \text{ENG})$.

III. Word's Extended Inclusion Axioms

(C) $\exists \mathcal{R}^c_p \in \mathcal{R}^c, \forall w_i, w_j \in W, \mathcal{R}^c_p(w_i, w_j) \wedge ExtInclude(w_i, w_j) \rightarrow \mathcal{R}^c_p$ is an *Abbreviation* relation;

(D) $\exists \mathcal{R}^c_p \in \mathcal{R}^c, \forall w_i, w_j \in W, \mathcal{R}^c_p(w_i, w_j) \wedge ExtInclude(w_j, w_i) \rightarrow \mathcal{R}^c_p$ is a *Full-Name* relation;

(C) is a sufficiency axiom of the *Abbreviation* relation, and (D) is sufficiency axiom of the *Full-Name* relation. $ExtInclude(x, y)$ is an extended inclusion relation, which represents the constitutions of y are from x.

IV. Co-referent Relation's Divergence and Convergence Axioms

The co-referent relation has two prominent features: *pointing-from* and *pointing-to* whose meaning are, for a co-referent relation $\mathcal{R}^c_p(w_i, w_j)$, that $\mathcal{R}^c_p$ points from w_i and points to w_j.

Definition 3: Given a target word w and a word set $W^o=\{w^o_1, \ldots, w^o_k\} \subset W$, for any $w^o_j \in W^o$ $(1 \leq j \leq k)$ such that $\mathcal{R}^c_p(w, w^o_j)$ holds, however, for any $w^{o'} \in W \backslash W^o$ that $\mathcal{R}^c_p(w, w^{o'})$ does not hold anymore, then the *pointing-from degree* of $\mathcal{R}^c_p$ on w is k, denoted as $P^O(\mathcal{R}^c_p, w)=k$.

Definition 4: Given a target word w and a word set $W^i=\{w^i_1, \ldots, w^i_h\} \subset W$, for any $w^i_j \in W^i$ $(1 \leq j \leq h)$ such that $\mathcal{R}^c_q(w^i_j, w)$ holds, however, for any $w^{i'} \in W \backslash W^i$ that $\mathcal{R}^c_q(w^{i'}, w)$ does not hold anymore, then the *pointing-to degree* of $\mathcal{R}^c_q$ on w is h, denoted as $P^I(\mathcal{R}^c_q, w)=h$.

Definition 5: Given a word set $W^O=\{w^o_1, \ldots, w^o_k\} \subset W$, for any $w^o_j \in W^O$ $(1 \leq j \leq k)$, there exists a word $w_f \in W$ such that $\mathcal{R}^c_p(w^o_j, w_f)$ holds, but there is no any word $w_g \in W \backslash W^O$

such that $\mathcal{R}_p(w_g, w_f)$ holds, then W^O is called the *pointing-from* set of $\mathcal{R}_p$ which means $\mathcal{R}_p$ points from any word of W^O.

Definition 6: Given a word set $W^I=\{w^i{}_1, \ldots, w^i{}_h\}\subset W$, for all $w^i{}_j \in W^I$ $(1\leq j\leq h)$, there exists $w_f \in W$ such that $\mathcal{R}_q(w_f, w^i{}_j)$ holds, but there is no any word $w_g \in W\backslash W^I$ such that $\mathcal{R}_q(w_f, w_g)$ holds, then W^I is called the *pointing-to* set of $\mathcal{R}_q$ which means $\mathcal{R}_q$ points to any word of W^I.

Given a co-referent relation $\mathcal{R}_p$ and its *pointing-from* set $W^O=\{w^o{}_1, \ldots, w^o{}_k\}$ and *pointing-to* set $W^I=\{w^i{}_1, \ldots, w^i{}_h\}$,

(E) The *divergence degree* of $\mathcal{R}_p$ is $Div(\mathcal{R}^{c}{}_p) = \underset{w^o_j \in W^O}{Max} P^O(\mathcal{R}^{c}{}_p, w^o_j)$, which means the *maximum pointing-from* degree of $\mathcal{R}_p$ on the set W^O.

(F) The *convergence degree* of $\mathcal{R}_p$ is $Cov(\mathcal{R}^{c}{}_p) = \underset{w^i_j \in W^I}{Max} P^I(\mathcal{R}^{c}{}_p, w^i_j)$, which means the *maximum pointing-to* degree of $\mathcal{R}_p$ on the set W^I.

V. Co-referent Token Words' Relation Axioms

Given two token words $\mathcal{R}_p$, $\mathcal{R}_q \in \mathcal{R}$, for any target word w_i, $w_j \in W$,

If $\mathcal{R}_p(w_i, w_j) \rightarrow \mathcal{R}_q(w_i, w_j)$ and $\mathcal{R}_q(w_i, w_j) \rightarrow \mathcal{R}_p(w_i, w_j)$ hold, then $\mathcal{R}_p$ and $\mathcal{R}_q$ are semantically equivalent, denoted as $\mathcal{R}_p \equiv \mathcal{R}_q$.

If $\mathcal{R}_p(w_i, w_j) \rightarrow \mathcal{R}_q(w_i, w_j)$ holds, but $\mathcal{R}_q(w_i, w_j) \rightarrow \mathcal{R}_p(w_i, w_j)$ does not hold, then $\mathcal{R}_p$ is contained by $\mathcal{R}_q$, denoted as $\mathcal{R}_p \Rightarrow \mathcal{R}_q$.

If $\mathcal{R}_p(w_i, w_j) \rightarrow \mathcal{R}_q(w_j, w_i)$ and $\mathcal{R}_q(w_j, w_i) \rightarrow \mathcal{R}_p(w_i, w_j)$ hold, then $\mathcal{R}_p$ and $\mathcal{R}_q$ are semantically reversible, denoted as $\mathcal{R}_p \equiv^{-1} \mathcal{R}_q$.

(G) $\exists \mathcal{R}_p, \mathcal{R}_q \in \mathcal{R}, \mathcal{R}_p \equiv \mathcal{R}_q; \forall w_i, w_j \in W, \mathcal{R}_p(w_i, w_j) \leftrightarrow \mathcal{R}_q(w_i, w_j)$;

(H) $\exists \mathcal{R}_p, \mathcal{R}_q \in \mathcal{R}, \mathcal{R}_p \Rightarrow \mathcal{R}_q; \forall w_i, w_j \in W, \mathcal{R}_p(w_i, w_j) \rightarrow \mathcal{R}_q(w_i, w_j)$;

(I) $\exists \mathcal{R}_p, \mathcal{R}_q \in \mathcal{R}, \mathcal{R}_p \equiv^{-1} \mathcal{R}_q; \forall w_i, w_j \in W, \mathcal{R}_p(w_i, w_j) \leftrightarrow \mathcal{R}_q(w_j, w_i)$.

(G) is a semantic equivalence axiom, (H) is a semantic implication axiom, and (I) is a semantic reverse axiom.

3.2 The Co-referent Relation Graph

If a target word is deemed as a vertex, and a co-referent relation as a directed edge, where a token word is a label of the directed edge, all candidate co-referent relation can be organized into a directed graph by their inter-links that is called co-referent relation graph.

Definition 7: Given a node set $V=\{\mathrm{v}_1, \ldots, \mathrm{v}_\mathrm{n}\}$, $v_i, v_j \in V(1\leq i, j\leq n)$, if $<v_i, v_j> \neq <v_j, v_i>$, then $<v_i, v_j>$ is called a ordered node tuple, abbreviated as ordered tuple.

Definition 8: A *semantic association graph* is a 4-tuple $SAG=(V, \mathcal{R}^a, E, f)$, where

- $V=\{v_1, \ldots, v_n\}$ is a set of target words;
- $\mathcal{R}^a=\{\mathcal{R}^a{}_1, \ldots, \mathcal{R}^a{}_t\}$ is a set of token words;
- $E=\{e_1, \ldots, e_m\}$ is a set of ordered relations;

- f is a mapping from E to the set of triples composed of the ordered tuples on V and $\mathcal{R}^a$.

Formally, for any directed edge $e_k \in E$ $(1 \leq k \leq m)$, there always exist two nodes v_i, $v_j \in V$ $(1 \leq i, j \leq n)$ and a relation token word $\mathcal{R}^a_h \in \mathcal{R}^a$ $(1 \leq h \leq t)$ such that $f(e_k) = <v_i, v_j, \mathcal{R}^a_h>$ holds. That is, e_k is a directed edge, labeled as $\mathcal{R}^a_h$, which points from v_i to v_j. v_i is the *beginning-node* of e_k, and v_j is the *ending-node* of e_k. According to e_k's direction, v_j is the *forward adjacency-node* of v_i. Reversely, v_i is the *backward adjacency-node* of v_j.

Definition 9: Given a semantic association graph $SAG=(V, \mathcal{R}^a, E, f)$, there exists $v \in V$, the number of edges whose ending-node is v is v's *in-degree*, and the number of edges whose beginning-node is v is v's *out-degree*.

Definition 10: Given a semantic association graph $SAG=(V, \mathcal{R}^a, E, f)$, for any $\mathcal{R}^a_h \in \mathcal{R}^a$ $(1 \leq h \leq t)$, if $\mathcal{R}^a_h$ is a co-referent token word, then the SAG is called a *co-referent relation graph*, denoted by $CRG=(V, \mathcal{R}^c, E, f)$, as illustrated in Fig.1.

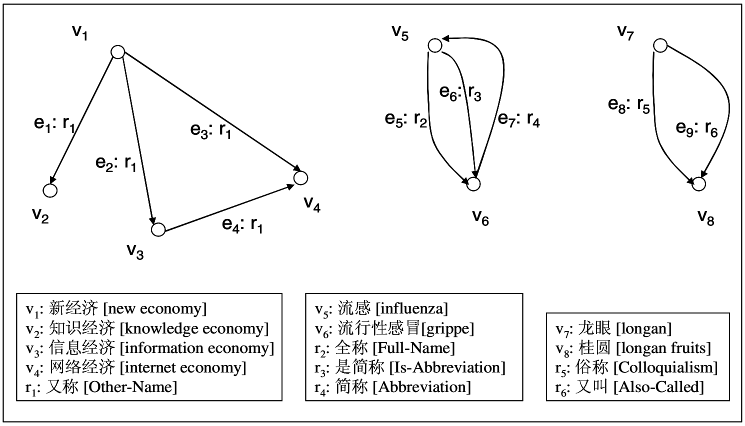

Fig. 1. A sample of co-referent relation graph

Definition 11: Given a co-referent relation graph $CRG=(V, \mathcal{R}^c, E, f)$, there exist a node $v \in V$ and a edge set $E^h \subseteq E$, for any edge $e_k \in E^h$, e_k has the same token word $\mathcal{R}^c_j$ $(\mathcal{R}^c_j \in \mathcal{R}^c)$,

If v is e_k's beginning node, then v's out-degree on $\mathcal{R}^c_j$ is called *homogeneous out-degree*, denoted as $OutDegree^H(v, \mathcal{R}^c_j)=|E^h|$, and e_k's ending node constitutes v's *homogeneously forward adjacency-node set*, denoted as $AdV^{P\text{-}H}(v, \mathcal{R}^c_j)$.

If v is e_k's ending node, then v's in-degree on $\mathcal{R}^c_j$ is called *homogeneous in-degree*, denoted as $InDegree^H(v, \mathcal{R}^c_j)=|E^h|$, and e_k's beginning node constitutes v's *homogeneously backward adjacency-node set*, denoted as $AdV^{N\text{-}H}(v, \mathcal{R}^c_j)$.

In Fig.1, for the node v_1 and the label r_1, $OutDegree^H(v_1, r_1)=3$, $AdV^{P\text{-}H}(v_1, r_1)=\{v_2, v_3, v_4\}$. For the node v_4 and the label r_1, $InDegree^H(v_4, r_1)=2$, $AdV^{N\text{-}H}(v_4, r_1)=\{v_1, v_3\}$.

3.3 Verification Strategies of Co-referent Relations

The basic verification idea is that the MFC constructs a co-referent relation graph after inputting all candidate relations, and then applies a series of axioms defined by the co-referent ontology to verify them. The graph is regulated continually till its structure does not change anymore. Finally correct candidates are outputted.

The first three types of axioms in the co-referent ontology are independent verification strategies which can verify a single candidate relation. The other two types are joint verification strategies that can only be used with the co-referent relation graph, and they can verify multiple candidates simultaneously.

Given a co-referent relation graph CRG=(V, $\mathcal{R}$, E, f):

There exist $w \in V$ and $\mathcal{R}_j \in \mathcal{R}$ such that w's homogeneous out-degree on $\mathcal{R}_j$ is $OutDegree^H(w, \mathcal{R}_j)$, and $\mathcal{R}_j$'s divergence degree is $Div(\mathcal{R}_j)$, or w's homogeneous in-degree on $\mathcal{R}_j$ is $InDegree^H(w, \mathcal{R}_j)$, and $\mathcal{R}_j$'s convergence degree is $Cov(\mathcal{R}_j)$,

1. If there is $OutDegree^H(w, \mathcal{R}_j) > Div(\mathcal{R}_j)$, then $\mathcal{R}_j(w, w_i)$ does not hold for all $w_i \in AdV^{P\text{-}H}(w, \mathcal{R}_j)$;
2. If there is $InDegree^H(w, \mathcal{R}_j) > Cov(\mathcal{R}_j)$, then $\mathcal{R}_j(w_i, w)$ does not hold for all $w_i \in AdV^{N\text{-}H}(w, \mathcal{R}_j)$.

There exist e_g, $e_h \in E$ and $\mathcal{R}_p$, $\mathcal{R}_q \in \mathcal{R}$ such that $f(e_g)$=<w_i, w_j, $\mathcal{R}_p$> and $f(e_h)$=<w_i, w_j, $\mathcal{R}_q$> hold,

3. If there is $\mathcal{R}_p \equiv \mathcal{R}_q$ or $\mathcal{R}_p \Rightarrow \mathcal{R}_q$, then $\mathcal{R}_p(w_i, w_j)$ and $\mathcal{R}_q(w_i, w_j)$ hold, otherwise, $\mathcal{R}_p(w_i, w_j)$ and $\mathcal{R}_q$ (w_i, w_j) do not hold.

There exist e_g, $e_h \in E$ and $\mathcal{R}_p$, $\mathcal{R}_q \in \mathcal{R}$ such that $f(e_g)$=<w_i, w_j, $\mathcal{R}_p$> and $f(e_h)$=<w_j, w_i, $\mathcal{R}_q$> hold,

4. If there is $\mathcal{R}_p \equiv^{-1} \mathcal{R}_q$, then $\mathcal{R}_p(w_i, w_j)$ and $\mathcal{R}_q(w_j, w_i)$ hold, otherwise, $\mathcal{R}_p(w_i, w_j)$ and $\mathcal{R}_q(w_j, w_i)$ do not hold.

In Fig.1, token words r_2 and r_3 have $r_2 \equiv r_3$, so $r_2(v_5, v_6)$ and $r_3(v_5, v_6)$ are both correct. Because token words r_5 and r_6 have $r_5 \Rightarrow r_6$, $r_5(v_7, v_8)$ and $r_6(v_7, v_8)$ are also correct. Both $r_3(v_5, v_6)$ and $r_4(v_6, v_5)$ are correct because of $r_3 \equiv^{-1} r_4$.

Except for the above verification strategies, the MFC can verify undecided relations using some already verified ones. This is called reinforced verification.

4 Experimental Analysis

From the Chinese open corpora of *2.6G* bytes, we use the predefined co-referent relation patterns (represented by regular expressions composed of token words and target words) to get the co-referent corpora (not limit domain, theme and style) of *18.4M* bytes. Using the MFC method, we acquire more than *60* types of co-referent relations from it. In Fig.2, the left is a group of sentences matched with patterns, and the right is a group of acquired relations from these sentences.

In the extraction phase, we get *66293* pieces of candidate relations, among which the distribution distance gets less than *3/4* of candidates, and the other two strategies get more than *1/4* as its supplements that increase the recall rate of extraction (see Table 1).

The verification orderly executes preprocessing, individual, joint and reinforced verification. Different strategies have different verification capabilities (see Table 2).

南方的老百姓很早就有吃鱼头的习惯，而且大多是吃鳙鱼（又叫花鲢鱼，俗称胖头鱼）。 (The people in south China have habit of eating fish-head for long history. They often eat big-head carp (also called spotted silver carp, colloquialism is fat-head fish))	⇒	又叫(鳙鱼, 花鲢鱼) i.e. Also-Called (big-head carp, spotted silver carp) 俗称(鳙鱼, 胖头鱼) i.e. Colloquialism (big-head carp, fat-head fish)

光动力疗法（简称PDT，photodynamic therapy）。
(photodynamic therapy (abbreviated as PDT))
	⇒	简称(光动力疗法, PDT) i.e. Abbreviation (photodynamic therapy, PDT)
		英文名称(光动力疗法, photodynamic therapy) i.e. English-Name(光动力疗法, photodynamic therapy)

科学家们说，严重急性呼吸道综合征（SARS，即非典型肺炎）
(Scientists said Severe Acute Respiratory Syndrome (SARS, that is Atypical Pneumonia))
	⇒	英文名称(严重急性呼吸道综合征, SARS) i.e. English-Name (严重急性呼吸道综合征, SARS)
		又称(严重急性呼吸道综合征, 非典型肺炎) i.e. Other-Name (Severe Acute Respiratory Syndrome, Atypical Pneumonia)

目前中国市场对聚氯乙烯（俗称PVC） (At present Polyvinyl Chloride in Chinese market (colloquialism is PVC))	⇒	俗称(聚氯乙烯, PVC) i.e. Colloquialism (Polyvinyl Chloride, PVC)

而根据美国风险投资协会（National Venture Capital Association 简称NVCA）
(However, according to National Venture Capital Association (abbreviated as NVCA))
	⇒	英文名称(美国风险投资协会, National Venture Capital Association) i.e. English-Name (美国风险投资协会, National Venture Capital Association)
		简称(美国风险投资协会, NVCA) i.e. Abbreviation (National Venture Capital Association, NVCA)

通常所说的失眠症乃是非器质性失眠症的简称。 (The so-called insomnia is in fact the abbreviation of Non-organic insomnia.)	⇒	简称(失眠症, 非器质性失眠症) i.e. Abbreviation (insomnia, Non-organic insomnia)
“小灵通”的学名是无线市话，是固定电话业务的延伸。 (*Xiaolingtong*’s academic name is wireless city-phone, and is an extension of fixed phone business.)	⇒	学名(小灵通, 无线市话) i.e. Academic-Name (Xiaolingtong, wireless city-phone)
西安古称长安，先后有12个王朝在此建都。 (Xi’an was called Chang’an in ancient times, which had been the capital of twelve dynasties.)	⇒	古称(西安, 长安) i.e. Ancient-Name (Xi’an, Chang’an)
囊状淋巴管瘤好发于颈部，又称囊状水瘤。 (Cystic lymphangioma often occurs in the neck, which has a other-name of hygroma.)	⇒	又称(囊状淋巴管瘤, 囊状水瘤) i.e. Other-Name (Cystic lymphangioma, hygroma)
对联又称楹联、门联、对子。 (Antithetical couplet’s other-name is pillar couplet, gatepost couplet, or couplet.)	⇒	又称(对联, 楹联) i.e. Other-Name (antithetical couplet, pillar couplet) 又称(对联, 门联) i.e. Other-Name (antithetical couplet, gatepost couplet) 又称(对联, 对子) i.e. Other-Name (antithetical couplet, couplet)

Fig. 2. Some examples of co-referent sentences and relations

Table 1. Performance of extraction strategies

Extraction Strategy	Number of Relations	Ratio (%)
distribution distance	48723	73.5%
pattern homogeneity	4312	6.5%
coordinate distribution	13258	20.0%
total	66293	100%

Table 2. Performance of verification strategies

Verification Strategy	Number of Relations	R-R (%)	F-R (%)
pre-processing	66293	87.93%	12.07%
individual verification	58293	88.20%	11.80%
joint verification	51414	65.16%	34.84%
reinforced verification	33501	92.24%	7.76%

The retained-ratio (R-R) is the ratio of retained relations after verifying. The filtered-ratio (F-R) is the ratio of filtered relations after verifying.

Table 3 lists different verification strategies' precision and recall rate on the acquisition. The benchmark is *66293* candidate relations. Finally, the precision rate (P) reaches *82.43%*, and the recall rate (R) reaches *92.03%*, and the error verification (correct relations but filtered) have *2206* pieces.

Table 3. Precision and recall rates of verification strategies

Verification Strategy	P (%)	R (%)	Verification Errors
pre-processing	46.70%	98.35%	457
individual verification	51.62%	95.88%	682
joint verification	76.27%	92.32%	987
reinforced verification	82.43%	92.03%	80

5 Conclusions

This paper proposes a multi-feature constrained method (MFC) for acquiring co-referent relations from large-scale Chinese corpora. It also provides some valuable guidance for other types of relation acquisitions. Such guidance is that it should use relations' features as many as possible, which include target words' distributive features and token words' semantic features, and structural features of all candidate relations. However, the acquisition by the MFC is restricted to the patterns. In the future, we will apply a pattern leaning method so that the relation and pattern acquisition would be integrated into one process to acquire more co-referent relations.

References

1. Marti A. Hearst, Automatic Acquisition of Hyponyms from Large Text Corpora, 14th International Conference on Computational Linguistics (COLING 1992), August 23-28, (1992), 539-545
2. Scott Cederberg and DominicWiddows, Using LSA and Noun Coordination Information to Improve the Precision and Recall of Automatic Hyponymy Extraction. Conference on Natural Language Learning (CoNLL-2003), Edmonton, Canada, (2003), 111-118
3. Rosana Girju, Adriana Badulescu and Dan Moldovan, Learning Semantic Constraints for the Automatic Discovery of Part-Whole Relations. In Proceedings of HLT-NAACL 2003, Edmonton, May-June, (2003), 1-8
4. Matthew Berland and Eugene Charniak, Finding Parts in Very Large Corpora. In Proceedings of the 37th Annual Meeting of the Association for the Computational Linguistics (ACL-99), College Park, MD, (1999), 57-64
5. Alexander Maedche and Steffen Staab, Discovering Conceptual Relations from Text. Proceedings of the 14th European Conference on Artificial Intelligence (ECAI 2000), Berlin, Germany, August 20-25, (2000), 321-325
6. Lin D and Pantel P, Induction of Semantic Classes from Natural Language Text. Proceedings of SIGKDD-01, San Francisco, CA, USA, (2001), 317-322