Construction of Diachronic Ontologies from *People's Daily* of Fifty Years

Shaoda He[1], **Xiaojun Zou**[2], **Liumingjing Xiao**[1], **Junfeng Hu**[1,2,*]
[1]Peking University, Beijing, P. R. China.
[2]Key Laboratory of Computational Linguistics, Ministry of Education, Peking University, Beijing, P. R. China.
{hesd, zouxj, xlmj, hujf}@pku.edu.cn

Abstract

This paper presents an Ontology Learning From Text (OLFT) method follows the well-known OLFT cake layer framework. Based on the distributional similarity, the proposed method generates multi-level ontologies from comparatively small corpora with the aid of HITS algorithm. Currently, this method covers terms extraction, synonyms recognition, concepts discovery and concepts hierarchical clustering. Among them, both concepts discovery and concepts hierarchical clustering are aided by the HITS authority, which is obtained from the HITS algorithm by an iteratively recommended way. With this method, a set of diachronic ontologies is constructed for each year based on *People's Daily* corpora of fifty years (i.e., from 1947 to 1996). Preliminary experiments show that our algorithm outperforms the Google's RNN and *K*-means based algorithm in both concepts discovery and concepts hierarchical clustering.

Keywords: Ontology Learning From Text (OLFT), Diachronic ontologies, HITS algorithm

1. Introduction

Previous research showed that distributional similarity based method achieved a helpful result in word semantic variation and change analysis on a diachronic corpus in both overall trends and word-level characteristics (Zou et al., 2013). However, this word-level analysis suffered from the problem of data sparseness. It is widely accepted that ontologies can facilitate text understanding and automatic processing of textual resources. Moving from words to concepts not only mitigates data sparseness issues, but also promises appealing solutions to polysemy and homonymy. Thus this paper aims at designing an Ontology Learning From Text (OLFT) method and applying it to construct a set of diachronic ontologies from such a diachronic corpus (i.e., *People's Daily* corpus from 1947 to 1996). These diachronic ontologies could be meaningful Chinese language resource for computational linguistics, sociolinguistics and related areas as they are promisingly more robust in diachronic analysis such as word semantic variation and change, concepts evolution, topics tracking, etc.

The OLFT approach designed in this paper follows the well-known OLFT cake layer framework (Cimiano, 2006). We adopt a distributional similarity based method to discover semantically similar words, and then a HITS (Kleinberg, 1999) and *K*-means (MacQueen, 1967) based method is applied to cluster these similar words hierarchically and a multi-level ontology is then generated. The proposed OLFT approach proved more flexibility on comparatively small corpora as the corpus for each year is not enough and tends to be sparse in ontology learning task. According to Sowa[1], ontologies can be categorized into three types: formal, prototype-based and terminological ontologies. The ontologies constructed in this paper is prototype-based and each concept is presented by a synset.

The contribution of this work is three-fold: 1) A new method on ontology learning from unstructured text on comparatively small corpora; 2) Publicly available[2] diachronic ontologies constructed from *People's Daily* from 1947 to 1996; 3) A fresh perspective on diachronic analysis provided by diachronic ontologies for computational linguistics, sociolinguistics and related areas.

2. Ontologies construction methodology

The OLFT method designed in this paper follows the steps described in well-known OLFT cake layer framework (Cimiano, 2006). According to this methodological approach, an ontology is built bottom-up starting from words that composing a text. First, domain-relevant terms are extracted, representing *domain terminology*. Terms are then aggregated into classes of synonyms and subsequently into concepts. The latter are then organized into a hierarchy or taxonomy through the relations of hyponymy and thereafter placed in relation with each other by means of non-taxonomic semantic relations. Finally, a set of rules is defined by means of logical inferences. At present, our ontology learning method just includes the first four layers from the bottom and we refer these steps as terms extraction, synonyms recognition, concepts discovery and concepts hierarchical clustering respectively.

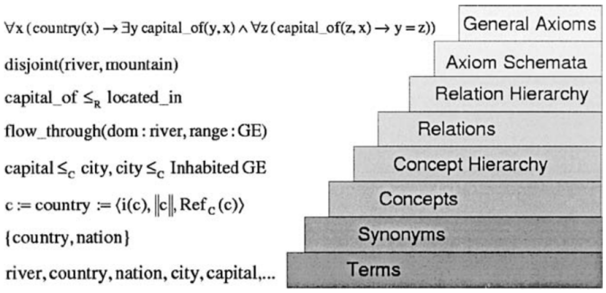

∀x (country(x) → ∃y capital_of(y, x) ∧ ∀z (capital_of(z, x) → y = z))
General Axioms
disjoint(river, mountain)
Axiom Schemata
capital_of ≤_R located_in
Relation Hierarchy
flow_through(dom : river, range : GE)
Relations
capital ≤_C city, city ≤_C Inhabited GE
Concept Hierarchy
c := country := ⟨i(c), ‖c‖, Ref_C (c)⟩
Concepts
{country, nation}
Synonyms
river, country, nation, city, capital,...
Terms

Figure 1: Ontology Learning Layer Cake (Cimiano, 2006)

*To whom all correspondence should be addressed.

[1] http://www.jfsowa.com/ontology/gloss.htm

[2] http://klcl.pku.edu.cn/clr/ontology/diachronic47-96.zip

2.1. Terms extraction

For simplicity, we segment the raw text of *People's Daily* and tag each of the words with a part of speech using Chinese Lexical Analysis System (ICTCLAS) (Zhang, 2002). All the words are taken into account except for stop words and low-frequency words.

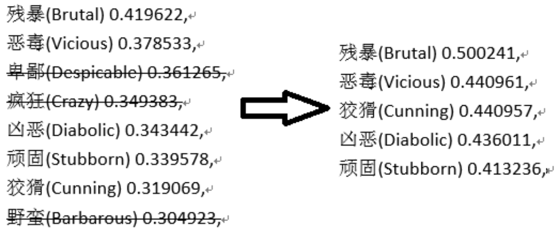
2.2. Synonyms recognition

Our method is based on the hypothesis of distributional similarity. Both lexical and syntactic contexts are considered in similarity computation. For lexical contexts, different window lengths are selected for terms with different parts of speech; For syntactic contexts, parts-of-speech of the neighboring words are considered. Thus, each term is represented by a vector associated with its distributional features. Each dimension of the vector is the PMI (Point-wise mutual information) of the corresponding feature. Afterwards, cosine similarity of each pair of terms is calculated in the subsequent synonyms recognition.

2.3. Concepts discovery

We adopt a HITS (Kleinberg, 1999) based algorithm to cluster terms into concepts. Given cosine similarity of each pair of terms, an empirical thresholds is set to retrieve a group of synonyms for each term. The term together with its synonyms can be viewed as initial concepts. Afterwards, the HITS algorithm is applied to enable terms in a initial concept to recommend each other iteratively. Then each term in the initial concept gains an authority value after the iterations convergence. A term may appear in several initial concepts. At present, our method ignores polysemy and homonymy, which means each term should be included in only its most related concept. The intimacy between the term and each of the concepts containing the term is represented by the distributional similarity weighted with the HITS authority. This intimacy is calculated and ranked. The term retains in the concept with highest intimacy and excluded from other concepts with lower intimacy.

Figure 2 illustrates how HITS algorithm is applied to exclude terms with lower intimacy to a certain concept and the terms with higher intimacy are retained in the concept. Terms (usually two terms) with highest authority are seen as semantic tags which represent major parts of the many aspects of the concept semantics. The top two terms in each concept are selected as the label of the concept, representing the meaning of this concept.



残暴(Brutal) 0.419622,↵
恶毒(Vicious) 0.378533,↵
~~卑鄙(Despicable) 0.361265,~~↵
~~疯狂(Crazy) 0.349383,~~↵
凶恶(Diabolic) 0.343442,↵
顽固(Stubborn) 0.339578,↵
狡猾(Cunning) 0.319069,↵
~~野蛮(Barbarous) 0.304923,~~↵

残暴(Brutal) 0.500241,↵
恶毒(Vicious) 0.440961,↵
狡猾(Cunning) 0.440957,↵
凶恶(Diabolic) 0.436011,↵
顽固(Stubborn) 0.413236,↵

Figure 2: An initial concept roughly generated is on the left, with some terms with lower intimacy excluded, the final concept with its composing terms is shown on the right, the intimacy between the term and the concept is also shown on the right of the term.

2.4. Concepts hierarchical clustering

In the following step, each concept is viewed as a term to do the HITS based clustering hierarchically. The hierarchical clustering of concepts is performed in a similar way to concept discovery described in 2.3. The slight difference is, when dealing with upper level of clusters, an iterative algorithm, *K*-means (MacQueen, 1967), is adopted to find a most appropriate larger cluster for a smaller one to be fixed into. Unlike the conventional *K*-means method, sub-center number *K* in our algorithm is not manually designated, but determined by similarity values between sub-cluster pairs and modifications of parameters. Given a fixed set of parameters, the ontology constructed in our algorithm is definite. Adequate iterations of *K*-means guarantee that, when a new level of clusters are merged, each of them contains highly semantically associated sub-clusters.

The detail of concepts hierarchical clustering is shown in ALGORITHM 1. The inputs of the algorithm are the first level concepts aggregated in the former step (denoted as $Conception$), matrix with similarities for each pair of terms (denoted as M_0), required levels of hierarchical clustering (denoted as n) and iterations for K-means algorithm (denoted as m). The output is the hierarchical clusters (denoted as $Ontology$).

```
ALGORITHM 1: ONTOLOGY GENERATION
ONTOLOGY-CLUSTERING($Conception, M_0, n, m$)
    Use $Conception$ to initialize $Ontology_1$;
    for $level \leftarrow 2$ to $n$
        Calculate matrix $M_{level-1}$ of similarities between
            pairs of $level-1$ clusters;
        Generate initial cluster $Cluster_0$ according to $M_{level-1}$;
        for $iteration \leftarrow 1$ to $m$
            Apply HITS Algorithm to every item in $Cluster_0$;
            Adjust $Cluster_0$ to form $Cluster_1$ according to
                HITS authority values;
            $Cluster_0 \leftarrow Cluster_1$;
        Record $Cluster_0$ after the loops above as $Ontology_{level}$;
return $Ontology$;
```

2.5. Diachronic ontologies construction

By applying the above steps to diachronic corpus of *People's Daily* (i.e., from 1947 to 1996) of each year, the yearly diachronic ontologies are constructed. As words in different times may have different senses, the diachronic ontologies could be meaningful Chinese language resource for computational linguistics, sociolinguistics and related areas as they are promisingly more robust in diachronic analysis such as word semantic variation and change, concepts evolution, topics tracking, etc.

3. Evaluation

To verify the effectiveness of ontologies constructed through our method, we choose Google's RNN and *K*-means based concepts discovery and concepts hierarchical clustering algorithm (which is implemented in the open source word2vec project[3]) as the baseline. We adopt HIT IR-Lab Tongyici Cilin (Extended)[4] provided by Harbin Institute of Technology as the standard for evaluating the quality of concepts clustering, by computing distances of words in each of our trees when mapping to Cilin. Since Cilin and our corpora cover not exactly the same vocabularies, we ignore words which do not appear in at least one of the trees. Average distances are calculated. Provided one result is perfect, its average distances should be 0.

Since the baseline approach requires cluster number before computation, we give it our Level 0 and 1 cluster numbers respectively. We calculate the average distance and the variance of all word pairs in an ontology when mapping to Cilin for both methods. As is shown in Table 1 and 2, our method achieves a better performance since its average distances are obviously shorter than the baseline. Although the average distances of our method are shorter than those of the baseline, they are still relatively large. Because our ontology mainly focuses on semantically similar words and their changes through time while Cilin is a static ontology (tree) for synonyms.

Method	Average Distance	Variance
Baseline	4.384	1.310
Our Approach	2.685	2.149

Table 1: Evaluation of concepts discovery using word-pair average distance (14,314 clusters for both approach).

Method	Average Distance	Variance
Baseline	4.383	1.318
Our Approach	3.416	1.903

Table 2: Evaluation of concepts hierarchical clustering using word-pair average distance (6,642 clusters for baseline method).

4. Language resource description

The raw data of our ontology construction is *People's Daily* of fifty years (i.e., from 1947 to 1996). We have constructed a set of diachronic ontologies and they are publicly available online[5].

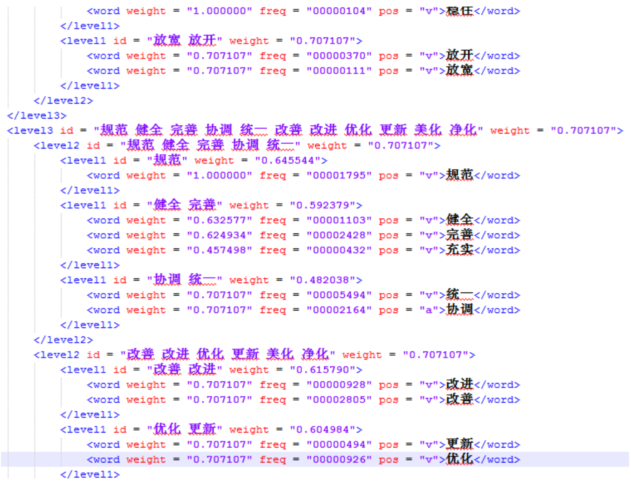
4.1. Annual diachronic ontologies

The ontology of each year contains 8 levels and we only consider words with frequencies not lower than 100. Numerals, punctuations, non-morpheme words, quantifiers and function words are excluded. Raw data sizes and vocabularies range from 26-130MB (with about 6M-12M words after segmentation) and 5,000-10,000 respectively. Take the year 1995's ontology with vocabularies of 9,991 as an example. Its nodes of levels from 0-8 are listed in Table 3.

Level	Nodes
0	9,991
1	5,985
2	2,765
3	1,290
4	600
5	251
6	96
7	33
8	12

Table 3: Nodes of levels from 0-8 in 1995's ontology (Nodes in Level 0 are words while in other levels are clusters).

Our ontologies are in XML format (as is shown in Figure 3). Each item in the lowest level denotes a term, and its attributes contain frequency in the year's corpus, its part-of-speech and its HITS authority weighted similarity in the cluster. Two terms with the highest values are selected as labels for each cluster. They can roughly indicate the senses of the cluster. For upper levels, labels are combined and they represent different aspects of a rather large cluster. The maximum number of words in a label is restricted to 20.



```
            <word weight = "1.000000" freq = "00000104" pos = "v">稳住</word>
        </level1>
        <level1 id = "放宽 放开" weight = "0.707107">
            <word weight = "0.707107" freq = "00000370" pos = "v">放开</word>
            <word weight = "0.707107" freq = "00000111" pos = "v">放宽</word>
        </level1>
    </level2>
</level3>
<level3 id = "规范 健全 完善 协调 统一 改善 改进 优化 更新 美化 净化" weight = "0.707107">
    <level2 id = "规范 健全 完善 协调 统一" weight = "0.707107">
        <level1 id = "规范" weight = "0.645544">
            <word weight = "1.000000" freq = "00001795" pos = "v">规范</word>
        </level1>
        <level1 id = "健全 完善" weight = "0.592379">
            <word weight = "0.632577" freq = "00001103" pos = "v">健全</word>
            <word weight = "0.624934" freq = "00002428" pos = "v">完善</word>
            <word weight = "0.457498" freq = "00000432" pos = "v">充实</word>
        </level1>
        <level1 id = "协调 统一" weight = "0.482038">
            <word weight = "0.707107" freq = "00005494" pos = "v">统一</word>
            <word weight = "0.707107" freq = "00002164" pos = "a">协调</word>
        </level1>
    </level2>
    <level2 id = "改善 改进 优化 更新 美化 净化" weight = "0.707107">
        <level1 id = "改善 改进" weight = "0.615790">
            <word weight = "0.707107" freq = "00000928" pos = "v">改进</word>
            <word weight = "0.707107" freq = "00002805" pos = "v">改善</word>
        </level1>
        <level1 id = "优化 更新" weight = "0.604984">
            <word weight = "0.707107" freq = "00000494" pos = "v">更新</word>
            <word weight = "0.707107" freq = "00000926" pos = "v">优化</word>
        </level1>
```

Figure 3: The sample XML format of our ontologies.

Our algorithm is able to produce relatively satisfactory result on a small corpus. For example, the corpus for 1977 is only 26MB (segmented text) and contains 4,269,940 words (including punctuations and all the other segments). The ontology is still semantically meaningful although fewer words are contained because of rather low word frequencies. Its nodes of levels from 0-8 are listed in Table 4.

The annual diachronic ontologies are suitable for researching on gradual semantic changes and concept revolution among consecutive years. However, the word frequencies are low and it is recommended to combine some consecutive years as a period to reduce data sparseness.

4.1.1. Diachronic ontologies of periods

The parameters set for period ontologies construction are similar to annual ones while word frequency is restricted to

[3] https://code.google.com/p/word2vec/

[4] http://ir.hit.edu.cn/demo/ltp/Sharing_Plan.htm

[5] http://klcl.pku.edu.cn/clr/ontology/diachronic47-96.zip

Level	Nodes
0	7,172
1	4,404
2	1,978
3	922
4	428
5	168
6	62
7	24
8	14

Table 4: Nodes of levels from 0-8 in 1977's ontology (Nodes in Level 0 are words while in other levels are clusters).

above 300 since the corpus for each period is relative larger. We manually divide the 53 years (from 1947 to 1999) into 8 periods according to major political events and corpora sizes. Evident political events considered are socialist transformation (before 1956), "3 years of natural disaster" (1959-1961), the Cultural Revolution (1966-1976), and etc. Corporal sizes of periods are around 400 Megabytes. The approximate data size are around 10,000 to 11,000 terms on the lowest level in each of the ontology. Table 5 lists the periods and their cumulative file sizes of segmented TXT format corpora. Table 6 shows the nodes of Level 0 to 8 in ontologies of the 8 periods.

Periods	Sizes of corpora (Megabytes)
1947-1954	406
1955-1960	440
1961-1967	425
1968-1976	409
1977-1983	405
1984-1988	370
1989-1994	427
1995-1999	381

Table 5: Manually divided periods and their respective raw data sizes.

5. Examples of diachronic analysis

By analysing synonyms in corpora of different eras, our method can reveal semantic changes of a term by comparing its neighboring terms or clusters.

Take the word "春风"(spring wind) as an example. Cilin relates it to other types of winds as is shown in Figure 4.

Our diachronic ontologies can show changes of word semantics through time. For example, in the era of the Cultural Revolution(1966-1976), the political meaning of "春风" (spring wind), positive changes of policies which benefit the people, is accentuated. So "春风" (spring wind) and "春雷" (spring thunder) are highly related with "喜讯" (good news) and "捷报" (report of success) in 1968-1976 corpora as is shown in Figure 5. During the year 1995-1999, the days of revolution are gone and the usage of "春风" (spring wind) mainly focuses on topics of weather. So we can see that in Figure 6 which partly shows the 1995-1999 result, words such as "风" (wind), "北风" (north wind), "雨" (rain), "雪" (snow) and "雾" (mist) are in its nearby clusters.

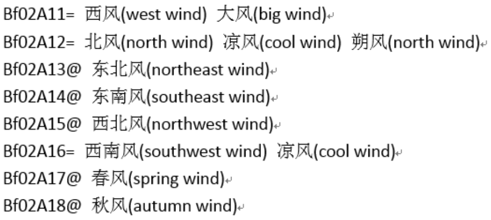

Bf02A11= 西风(west wind) 大风(big wind)
Bf02A12= 北风(north wind) 凉风(cool wind) 朔风(north wind)
Bf02A13@ 东北风(northeast wind)
Bf02A14@ 东南风(southeast wind)
Bf02A15@ 西北风(northwest wind)
Bf02A16= 西南风(southwest wind) 凉风(cool wind)
Bf02A17@ 春风(spring wind)
Bf02A18@ 秋风(autumn wind)

Figure 4: "春风" (spring wind) and its synonyms in Cilin. The left column shows the precise position of the lowest-level clusters in Cilin ontology (tree). These small clusters (in level 1) belong to the same larger cluster in level 2.

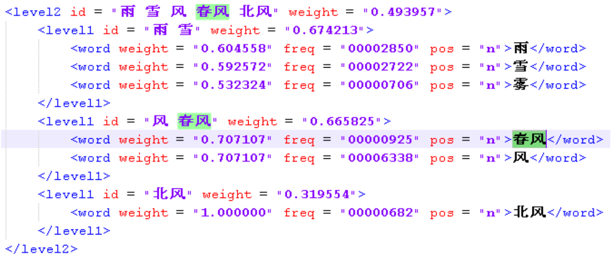

```
<level2 id = "雨 雪 风 春风 北风" weight = "0.493957">
    <level1 id = "雨 雪" weight = "0.674213">
        <word weight = "0.604558" freq = "00002850" pos = "n">雨</word>
        <word weight = "0.592572" freq = "00002722" pos = "n">雪</word>
        <word weight = "0.532324" freq = "00000706" pos = "n">雾</word>
    </level1>
    <level1 id = "风 春风" weight = "0.665825">
        <word weight = "0.707107" freq = "00000925" pos = "n">春风</word>
        <word weight = "0.707107" freq = "00006338" pos = "n">风</word>
    </level1>
    <level1 id = "北风" weight = "0.319554">
        <word weight = "1.000000" freq = "00000682" pos = "n">北风</word>
    </level1>
</level2>
```

Figure 5: Part of 1968-1976 ontology showing "春风" (spring wind) and its semantically similar words.

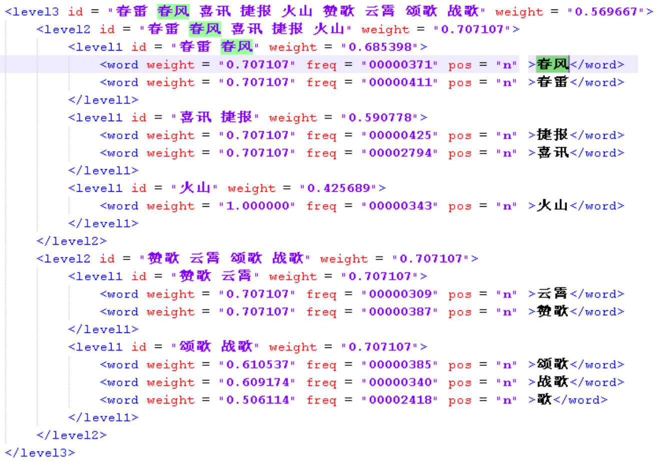

```
<level3 id = "春雷 春风 喜讯 捷报 火山 赞歌 云霄 颂歌 战歌" weight = "0.569667">
    <level2 id = "春雷 春风 喜讯 捷报 火山" weight = "0.707107">
        <level1 id = "春雷 春风" weight = "0.685398">
            <word weight = "0.707107" freq = "00000371" pos = "n" >春风</word>
            <word weight = "0.707107" freq = "00000411" pos = "n" >春雷</word>
        </level1>
        <level1 id = "喜讯 捷报" weight = "0.590778">
            <word weight = "0.707107" freq = "00000425" pos = "n" >捷报</word>
            <word weight = "0.707107" freq = "00002794" pos = "n" >喜讯</word>
        </level1>
        <level1 id = "火山" weight = "0.425689">
            <word weight = "1.000000" freq = "00000343" pos = "n" >火山</word>
        </level1>
    </level2>
    <level2 id = "赞歌 云霄 颂歌 战歌" weight = "0.707107">
        <level1 id = "赞歌 云霄" weight = "0.707107">
            <word weight = "0.707107" freq = "00000309" pos = "n" >云霄</word>
            <word weight = "0.707107" freq = "00000387" pos = "n" >赞歌</word>
        </level1>
        <level1 id = "颂歌 战歌" weight = "0.707107">
            <word weight = "0.610537" freq = "00000385" pos = "n" >颂歌</word>
            <word weight = "0.609174" freq = "00000340" pos = "n" >战歌</word>
            <word weight = "0.506114" freq = "00002418" pos = "n" >歌</word>
        </level1>
    </level2>
</level3>
```

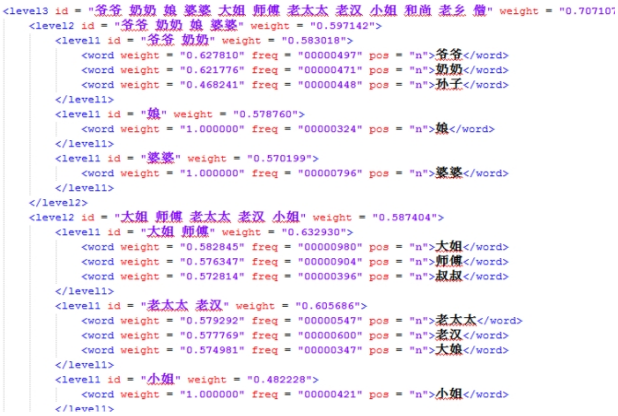
Figure 6: Part of 1995-1999 ontology showing "春风" (spring wind) and its semantically similar words.

Semantic changes may lead to polysemy. Figure 7 and 8 indicates the semantic changes of "小姐" (miss, young lady) according to ontologies of in 1984-1988, 1989-1994 and 1995-1999. As is shown in the figure, "小姐" mainly refers to lady or attractive young woman in the corpora of the 1980s. While in the early 1990s, it mostly means waitress (e.g. restaurant waitress or ritual girl) in the service industry since China's economy was expanding at an

Periods	0	1	2	3	4	5	6	7	8
1947-1954	10,334	4,469	2,124	1,007	441	184	85	28	9
1955-1960	10,878	4,774	2,210	1,082	506	221	73	26	11
1961-1967	10,915	5,093	2,368	1,134	534	225	83	24	9
1968-1976	9,951	5,041	2,319	1,102	491	214	80	28	12
1977-1983	11,619	5,677	2,697	1,295	591	257	106	42	14
1984-1988	11,443	5,507	2,563	1,225	545	222	89	33	12
1989-1994	13,097	6,055	2,690	1,277	576	242	96	34	15
1995-1999	11,725	5,904	2,702	1,269	557	218	85	33	14

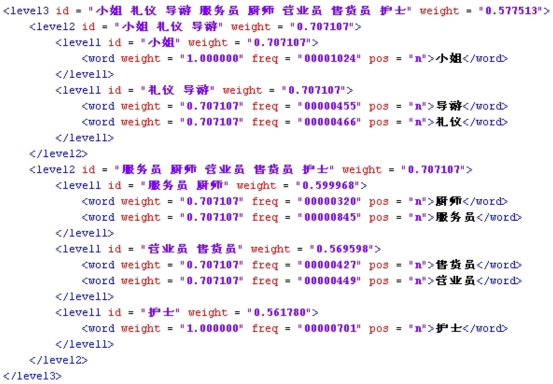
Table 6: Nodes of Level 0 to 8 in ontologies of the 8 periods.

amazing speed after the opening and reform policy. In the late 1990s, the word implies other aspects and it may be on the way to develop polysemy again. With “老板”(boss) in the same cluster, “小姐” might have gained meanings of female secretary or prostitute.



```
<level3 id = "爷爷 奶奶 娘 婆婆 大姐 师傅 老太太 老汉 小姐 和尚 老乡 僧" weight = "0.707107
    <level2 id = "爷爷 奶奶 娘 婆婆" weight = "0.597142">
        <level1 id = "爷爷 奶奶" weight = "0.583018">
            <word weight = "0.627810" freq = "00000497" pos = "n">爷爷</word>
            <word weight = "0.621776" freq = "00000471" pos = "n">奶奶</word>
            <word weight = "0.468241" freq = "00000448" pos = "n">孙子</word>
        </level1>
        <level1 id = "娘" weight = "0.578760">
            <word weight = "1.000000" freq = "00000324" pos = "n">娘</word>
        </level1>
        <level1 id = "婆婆" weight = "0.570199">
            <word weight = "1.000000" freq = "00000796" pos = "n">婆婆</word>
        </level1>
    </level2>
    <level2 id = "大姐 师傅 老太太 老汉 小姐" weight = "0.587404">
        <level1 id = "大姐 师傅" weight = "0.632930">
            <word weight = "0.582845" freq = "00000980" pos = "n">大姐</word>
            <word weight = "0.576347" freq = "00000904" pos = "n">师傅</word>
            <word weight = "0.572814" freq = "00000396" pos = "n">叔叔</word>
        </level1>
        <level1 id = "老太太 老汉" weight = "0.605686">
            <word weight = "0.579292" freq = "00000547" pos = "n">老太太</word>
            <word weight = "0.577769" freq = "00000600" pos = "n">老汉</word>
            <word weight = "0.574981" freq = "00000347" pos = "n">大娘</word>
        </level1>
        <level1 id = "小姐" weight = "0.482228">
            <word weight = "1.000000" freq = "00000421" pos = "n">小姐</word>
        </level1>
```

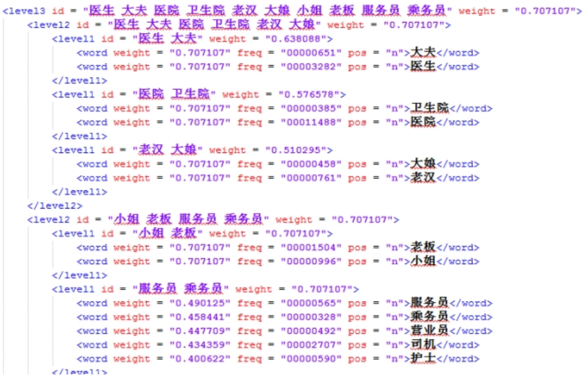
Figure 7: Part of 1984-1988 ontology showing “小姐” (miss, young lady) and its semantically similar words.



```
<level3 id = "小姐 礼仪 导游 服务员 厨师 营业员 售货员 护士" weight = "0.577513">
    <level2 id = "小姐 礼仪 导游" weight = "0.707107">
        <level1 id = "小姐" weight = "0.707107">
            <word weight = "1.000000" freq = "00001024" pos = "n">小姐</word>
        </level1>
        <level1 id = "礼仪 导游" weight = "0.707107">
            <word weight = "0.707107" freq = "00000455" pos = "n">导游</word>
            <word weight = "0.707107" freq = "00000466" pos = "n">礼仪</word>
        </level1>
    </level2>
    <level2 id = "服务员 厨师 营业员 售货员 护士" weight = "0.707107">
        <level1 id = "服务员 厨师" weight = "0.599968">
            <word weight = "0.707107" freq = "00000320" pos = "n">厨师</word>
            <word weight = "0.707107" freq = "00000845" pos = "n">服务员</word>
        </level1>
        <level1 id = "营业员 售货员" weight = "0.569598">
            <word weight = "0.707107" freq = "00000427" pos = "n">售货员</word>
            <word weight = "0.707107" freq = "00000449" pos = "n">营业员</word>
        </level1>
        <level1 id = "护士" weight = "0.561780">
            <word weight = "1.000000" freq = "00000701" pos = "n">护士</word>
        </level1>
    </level2>
</level3>
```

Figure 8: Part of 1989-1994 ontology showing “小姐” (miss, young lady) and its semantically similar words.

Nevertheless, the changes of synonyms or neighboring clusters of a term does not always denoting semantic changes of the term. Another exception is that a new topic may appear in a specific era and the similar terms for the topic emerge and change through ontologies of different years or periods. For example, during 1977-1983, “考试” (examination) and “高考” (National College Entrance Examination, NCEE) are highly similar. In the meantime, new terms such as “函授” (teaching by correspondence) and “自学” (self-study) appear due to new phenomena in education. However, it does not necessarily mean that the semantics of “考试” (examination) have evident changes. Because in the early 1980s, “高考” (NCEE), “函授” (teaching by correspondence) and “自学考试” (self-study examination) became hot topics after the Cultural Revolution, a dark age when learning was abandoned and condemned. And “考试” (examination) is highly related to the topic.



```
<level3 id = "医生 大夫 医院 卫生院 老汉 大娘 小姐 老板 服务员 乘务员" weight = "0.707107">
    <level2 id = "医生 大夫 医院 卫生院 老汉 大娘" weight = "0.707107">
        <level1 id = "医生 大夫" weight = "0.638088">
            <word weight = "0.707107" freq = "00000651" pos = "n">大夫</word>
            <word weight = "0.707107" freq = "00003282" pos = "n">医生</word>
        </level1>
        <level1 id = "医院 卫生院" weight = "0.576578">
            <word weight = "0.707107" freq = "00000385" pos = "n">卫生院</word>
            <word weight = "0.707107" freq = "00011488" pos = "n">医院</word>
        </level1>
        <level1 id = "老汉 大娘" weight = "0.510295">
            <word weight = "0.707107" freq = "00000458" pos = "n">大娘</word>
            <word weight = "0.707107" freq = "00000761" pos = "n">老汉</word>
        </level1>
    </level2>
    <level2 id = "小姐 老板 服务员 乘务员" weight = "0.707107">
        <level1 id = "小姐 老板" weight = "0.707107">
            <word weight = "0.707107" freq = "00001504" pos = "n">老板</word>
            <word weight = "0.707107" freq = "00000996" pos = "n">小姐</word>
        </level1>
        <level1 id = "服务员 乘务员" weight = "0.707107">
            <word weight = "0.490125" freq = "00000565" pos = "n">服务员</word>
            <word weight = "0.458441" freq = "00000328" pos = "n">乘务员</word>
            <word weight = "0.447709" freq = "00000492" pos = "n">营业员</word>
            <word weight = "0.434359" freq = "00002707" pos = "n">司机</word>
            <word weight = "0.400622" freq = "00000590" pos = "n">护士</word>
        </level1>
```

Figure 9: Part of 1995-1999 ontology showing “小姐” (miss, young lady)and its semantically similar words.

6. Conclusions and Future Work

This paper proposes a HITS based ontology learning algorithm from unstructured Chinese text and presents a set of diachronic ontologies constructed from *People's Daily* corpora of fifty years (i.e., from 1947 to 1996). Preliminary experiments showed that the proposed method outperforms Google's RNN and *K*-means based algorithm in both concepts discovery and concepts hierarchical clustering for small-scale and incremental corpora. The diachronic ontologies could be meaningful Chinese language resource for computational linguistics, sociolinguistics and related areas as they are promisingly more robust in diachronic analysis such as word semantic variation and change, concepts evolution, topics tracking, etc.

Further researches may include the following aspects. Firstly, polysemy and homonymy should be considered. Secondly, there are other important aspects of ontology learning, such as relationship and axiom schema learning and etc. And how to compare and merge similar parts of ontologies in different eras is also a tough problem.

7. Acknowledgements

This research is supported by the National Natural Science Foundation of China (Grant No. M1321005).

8. References

Cimiano P. 2006. Ontology learning and population from text:algorithms, evaluation and applications. Springer.

Fano, R.M. 1961. Transmission of Information: A Statistical Theory of Communications. In *American Journal of Physics*, pages 793-794.

Harris, Z.S. 1954. Distributional Structure. In *Word*, pages 146-162.

Hearst, M.A. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proc. Int'l Conf. Computational Linguistics (COLING)*, pages 539-545.

Kleinberg, J. M. 1999. Authoritative sources in a hyperlinked environment. In *Journal of the ACM(JACM)*, pages 604-632.

MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, pages 281-297.

Mikolov, T., Yih, W. T., & Zweig, G. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of NAACL-HLT*, pages 746-751.

Zhang H.P., Liu Q., Cheng X.Q. , Yu H.K. 2003. Chinese lexical analysis using hierarchical hidden markov model. In *Proceedings of the second SIGHAN workshop on Chinese language processing*, pages 63-70.

Zou, X., Sun, N., Zhang, H., & Hu, J. 2013. Diachronic Corpus Based Word Semantic Variation and Change Mining. In *Language Processing and Intelligent Information Systems*, pages 145-150.