Construction of Diachronic Ontologies from People’s Daily of Fifty Years

Shaoda He¹, Xiaojun Zou², Liumingjing Xiao¹, Junfeng Hu¹,²,∗
¹Peking University, Beijing, P. R. China.
²Key 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). 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 classes of synonyms and subsequently into concepts. The OLFT method designed in this paper follows the step-s 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 hierarchies or taxonomies 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.

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

2To whom all correspondence should be addressed.
1http://www.jfsowa.com/ontology/gloss.htm
2http://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 threshold 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 an 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 concepts, and low-frequency words.

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.

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 M0), 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

1. ONTOLOGY-CLUSTERING(Conception, M0, n, m)
2. Use Conception to initialize Ontology;
3. for level ← 2 to n
4. Calculate matrix M_{level−1} of similarities between pairs of level − 1 clusters;
5. Generate initial cluster Cluster0 according to M_{level−1};
6. for iteration ← 1 to m
7. Apply HITS Algorithm to every item in Cluster0;
8. Adjust Cluster0 to form Cluster1 according to HITS authority values;
9. Cluster0 ← Cluster1;
10. Record Cluster0 after the loops above as Ontology_{level};
11. 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

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.
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 approaches).

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

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

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

| 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                         |

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.

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
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.

![Figure 7](part-of-1984-1988-ontology-showing-%E5%B0%8F%E5%A5%A5.png)  
Figure 7: Part of 1984-1988 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 top-ic 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 e-ducation. 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.

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.