Diachronic Deviation Features in Continuous Space Word Representations

Ni Sun[1,2], Tongfei Chen[1], Liumingjing Xiao[1], Junfeng Hu[1,2⋆]

[1] School of Electronics Engineering & Computer Science,
Peking University, Beijing, P. R. China
[2] Key Laboratory of Computational Linguistics (Ministry of Education),
{sn96,ctf,hujf}@pku.edu.cn
xlmj531@163.com

Abstract. In distributed word representation, each word is represented as a unique point in the vector space. This paper extends this to a diachronic setting, where multiple word embeddings are generated with corpora in different time periods. These multiple embeddings can be mapped to a single target space via a linear transformation. In this target space each word is thus represented as a distribution. The deviation features of this distribution can reflect the semantic variation of words through different time periods. Experiments show that word groups with similar deviation features can indicate the hot topics in different ages. And the frequency change of these word groups can be used to detect the age of peak celebrity of the topics in the history.

Keywords: Lexical semantics, diachronic corpora, semantic distribution, hot topics

1 Introduction

Representation of words as dense, real-valued vectors can be trained via a neural network language model[1, 7]. It has been shown that these distributed representations of words can be used to improve the performance of many NLP systems[3].

However, despite such models' successful application, most of these models do not consider the concept of diachronicity, i.e. the change of the semantics of the words through different time periods is not taken into account. In this paper, we devised a feature vector that represents the semantic variation of a word in a diachronic corpus.

Mikolov et al. [8] demonstrated that it is possible to produce a linear projection between vector spaces of words that represent different languages. We adopt the idea to vector spaces of words that represents texts of different time periods. Word embeddings learned from different time periods are projected to the same vector space (target space).

⋆ Corresponding author

For a specific word, a multinomial Gaussian distribution is defined to fit all the projections in the target space. The deviation feature vector drawn from this distribution reflects the stability of the semantic of this word in a diachronic corpus. Experiments showed that these deviation features can be used in mining hot topics in different ages.

The rest of this paper is organized as follows. Section 2 elaborates the details of the mapping of different vector spaces. Section 3 illustrates our diachronic deviation feature vector, and its use in topic clustering. In Section 4, evaluations are presented to illustrate the effectiveness of the diachronic deviation feature, and the application of the topic cluster is also presented. The final section concludes this paper and discusses possible future work.

2 Linear Projection between Spaces

Given a diachronic corpus, we seek to split it into diachronic sections, under the assumption that each section is synchronic. For each section a distributed representation of words is trained via the method proposed by Mikolov et al. [7]. Then, by analogy of the method proposed by Mikolov et al. [8], a linear projection is built to transform all these vector spaces into a target space. It serves as a transformation that transforms all spaces to a uniform one. The symbols used in this section are described below:

Table 1. Symbols

Symbol	Definition
$\mathbf{x}(w, s)$	Vector representation of word w in slice s
$\mathbf{T}_{st}$	Transformation matrix from slice s to t

2.1 Splitting the Diachronic Corpus

A sliding-window based scheme is used to split the diachronic corpus. The splitting is dependent on two variables: window size and window increment. For example, in Figure 1, the window size is 5 years and the window increment is 1 year.

Using this method a diachronic corpus (year 1947 to 1996 in Figure 1) can be split into multiple (46 here) segments, in which each segment contains a time-consecutive portion of the original corpus, while large enough to train a reliable distributed representation of words. Additionally, overlapping of the slices, instead of using disjoint slices, produces more samples.

2.2 Training the Linear Projection

In order to observe the representation of words over different time periods, a liner transformation is used here to project all the different vector spaces into a target space.

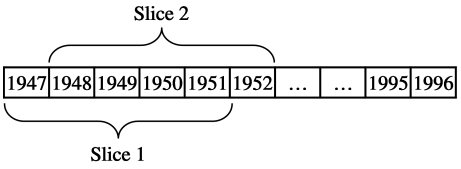

Fig. 1. Sliding window scheme.

In our experiments, the *target space* is generated by training the entire diachronic corpus using the neural network model. We seek to transform each vector space of a time period (*source space*) to the target space. This problem is formulated below:

Problem description: Given a set of words and their associated vector representations in two time periods

$$(\mathbf{x}(w,s),\mathbf{x}(w,t))_{i=1}^{n} \ , \tag{1}$$

find a transformation matrix $\mathbf{T}_{st}$ such that $\mathbf{T}_{st}\mathbf{x}(w,s)$ approximates $\mathbf{x}(w,t)$.

In practice, $\mathbf{T}_{st}$ can be learned by the following optimization problem:

$$\min_{\mathbf{T}_{st}} \sum_{i=1}^{n} \left\|\mathbf{T}_{st}\mathbf{x}(w_i,s) - \mathbf{x}(w_i,t)\right\|^2 . \tag{2}$$

This is solved using batch gradient descent.

After the transformation matrix $\mathbf{T}_{st}$ is trained, we can transform every vector in the source space to the target space by computing

$$\hat{\mathbf{x}}(w,t) = \mathbf{T}_{st}\mathbf{x}(w,s) \ . \tag{3}$$

Despite its simplicity, this linear transformation worked well between different languages Mikolov et al. [8], and it performed effectively in our experiments as well.

2.3 Generating the Training Data Set

One of the key problems that influence the result of the transformation matrix is the proper choice of the training set. In this section we focus on how to build an appropriate training set for the optimization problem stated above, i.e. a set of words with their associated vector representations in two spaces.

In this task, a set of words whose semantics are stable over time periods is desired. To avoid overfitting, the size of this set should be relatively small. In our experiments 100 words are selected for both Chinese and English corpus. First, we build this training set beginning from a randomly selected set. Initial transformation matrices are trained from this initial training set with the target space. Then word whose variances of the error between the actual representation

Algorithm 1. GenerateTrainingDataSet

Input: Set of words that occurred in all slices W
Output: Training data set W'

begin
  Randomly select k words as the initial set
  for i from 1 to T
    Train the transformation matrix $\mathbf{T}_{i0}$ for slice i
  end
  for $w \in W$
    for i from 1 to T
      $\mathrm{d(w,i)} = \|\mathbf{x}(w,0) - T_{i0}\mathbf{x}(w,i)\|$
    end
    $v(w) = var(w, \cdot)$
  end
  Return the top n words with the smallest $v(w)$
end

and the predicted representation are lowest is selected. In Algorithm 1, the target space is numbered 0.

Table 2 presents the generated training set of words (top 25 shown here) on two diachronic corpora: one in Chinese (*People's Daily*) and one in English (*New York Times*).

Table 2. Generated training data set for building the linear transformation matrix. (top 24 shown here)

Chinese (*People's Daily*)		English (*New York Times*)	
不仅(not only)	而且(and)	while	usually
例如(such as)	但是(but)	and	which
当时(at that time)	以前(previously)	now	place
还要(still)	地方(place)	although	industry
就是(exactly)	因为(because)	called	but
因此(hence)	这个(this)	similarly	still
这些(these)	等等(and so on)	also	presumably
同时(meanwhile)	原来(it turns out)	quietly	whereas
所以(therefore)	至于(as for)	mostly	suddenly
并且(as well as)	经过(after)	supposedly	apparently
除了(except)	虽然(although)	continually	with
只是(just)	当然(of course)	fully	however

From the generated words we could see that the training set generated mostly contains conjunctions and adverbs or some common concepts in both Chinese and English. The semantics of these words are mostly stable in diachronic cor-

pora, thus serving as a good training set for the training of the linear transformation model over time periods.

3 Diachronic Deviation Feature

From the linear transformation procedure described above, for each word w, we have a set of vector representations drawn from different time periods all projected to the same space, namely

$$S(w) = \{\mathbf{T}_{s0}\mathbf{x}(w, s)\}_{s=1,\cdots,T}. \tag{4}$$

This set of vectors $S(w)$ is fit to a multidimensional Gaussian distribution, namely the *diachronic representation distribution* $N(\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w)$.

From the word representation model it can be assumed that the dimensions of the representation space are mutually independent. Thus, the covariance matrix $\boldsymbol{\Sigma}_w$ is reduced to a diagonal matrix. We define the deviation features as the variance of each dimension:

$$\boldsymbol{\varphi}_w = (\Sigma_{ii})_{i=1,\cdots,D}\ , \tag{5}$$

where D is the dimensionality of the vector space. A min-max scaling is performed on these features:

$$\tilde{\varphi}_w^{(i)} = \frac{\varphi_w^{(i)} - \min \varphi_{(\cdot)}^{(i)}}{\max \varphi_{(\cdot)}^{(i)} - \min \varphi_{(\cdot)}^{(i)}}\ . \tag{6}$$

Here $\varphi^{(i)}$ denotes the ith dimension of vector $\boldsymbol{\varphi}$. The scaled vector $\tilde{\boldsymbol{\varphi}}$ is our diachronic deviation feature vector.

4 Evaluation

In this section, we demonstrate that clustering result of words using the diachronic deviation feature vectors is correlated with time-specific topics.

4.1 Corpus and Experiment Settings

Experiments are conducted on one Chinese real-word corpora: *People's Daily* of 50 years (from 1947 to 1996). ICTCLAS[9] is applied to segment the raw text. Window size is 5 years and window increment is set to be 1 year. Each text slice contains approximately 30 million words.

For each text slice, the vector representation is learned using the method by Mikolov et al. [7]. The dimension of the vectors is set to be 50. The target vector space is trained by the entire corpus. The words used for clustering are words that are prevalent in every time slice; i.e. the number of occurrence of a specific word in any time slice is greater than a minimum threshold. The actual number of words for clustering is approximately 10000.

4.2 Clustering

A diachronic deviation feature vector is generated for each word using the method described in Section 3. We use the cosine similarity measure as the similarity measure between the feature vectors. And we used the hierarchical word clustering scheme described by He et al. [4]. It uses the hyper-link induced topic search algorithm [5] to produce clusters of words. The number of clusters is determined after the completion of the algorithm, thus it is not necessary for the users to specify the number of clusters. The algorithm is shown below.

Algorithm 2. HierarchicalClustering

Input: Level-1 concept set C, level n, Similarity matrix of words M_0
Output: Hierarchical clustering tree H

```
1.  begin
2.    for l from 2 to n
3.      M_l ← Similarity matrix in level-(l − 1) concepts
4.      Perform initial clustering according to M_l
5.      while maximum iteration count not reached
6.        Run HITS on each concept to get authority score a
7.        Adjust the clustering result according to a
8.      end
9.      Write the clustering result of level-l to H
10.   end
11. end
```

We present several clusters produced by our method in Table 3. It can be seen that the produced clusters are largely correlated with topics instead of synonyms. Namely, The words in the same cluster have a tendency to occur in a specific time period, and they are correlated with a same hot topic in that period.

4.3 Case Study on *People's Daily*

In this section we present the experimental results on the Chinese corpus People's Daily (from 1947 to 1996). Table 4 is a cluster generated using the method described in Section 4.2.

It can be seen that the words in the cluster are closely related to topics concerning with exploiting class and revolution. The normalized frequency of these terms in the entire diachronic corpus is illustrated in Figure 2, where the frequency is divided by the total frequency of all the terms of each year. An abrupt change of frequency can be noticed around 1967.

Frequency of these terms is not uniformly distributed with respect to time; the terms have a tendency to occur in a specific time period. Then we choose the

Table 3. Examples of clusters produced by the similarity of deviation features.

Higher education	师范大学(normal university), 清华大学(Tsinghua University), 北京大学(Peking University), 学院(College), 工学院(College of Engineering), 医学院(College of Medicine), 师范学院(Normal College), 研究室(Laboratory), 院校(Institute)
Cities	西安市(Xi'an), 杭州市(Hangzhou), 沈阳市(Shenyang), 广州市(Guangzhou), 武汉市(Wuhan), 长春市(Changchun), 南京市(Nanjing), 天津市(Tianjin), 上海市(Shanghai), 长沙市(Changsha), 包头(Baotou), 北京市(Beijing), 重庆市(Chongqing)
Meteorological phenomena	洪水(flood), 山洪(mountain torrents), 雨(rain), 风沙(sandwind), 霜(frost), 雹(hail), 猛涨(surge), 风暴(storm)
Kinships	哥哥(elder brother), 姐姐(elder sister), 女儿(daughter), 妻子(wife), 父亲(father), 弟弟(younger brother), 母亲(mother), 丈夫(husband), 家里(at home), 儿子(son), 爱人(lover), 老乡(folks), 生病(sick), 妹妹(younger sister), 照料(taking care of), 孩子(child), 叔叔(uncle), 娃娃(kid), 孙子(grandchild), 邻居(neighbor)

Table 4. A cluster of words generated from *People's Daily*

统治(rule), 统治者(ruler), 封建(feudal), 官僚(bureaucrat), 专制(despotism), 势力(force), 统治阶级(ruling class), 剥削阶级(exploiting class), 推翻(overturn), 剥削(exploit), 封建主义(feudalism), 资产阶级(bourgeoisie), 右派(the Right), 国民党(Kuomingtang), 资本家(capitalist), 残余(remnant), 左派(the Left), 农奴(serf), 执政(be in power), 地主(landlord), 压迫(oppress), 腐败(corrupt), 改组(reorganize), 殖民(colonization), 派别(faction), 反动(reaction), 垄断(monopoly), 瓦解(disintegration)

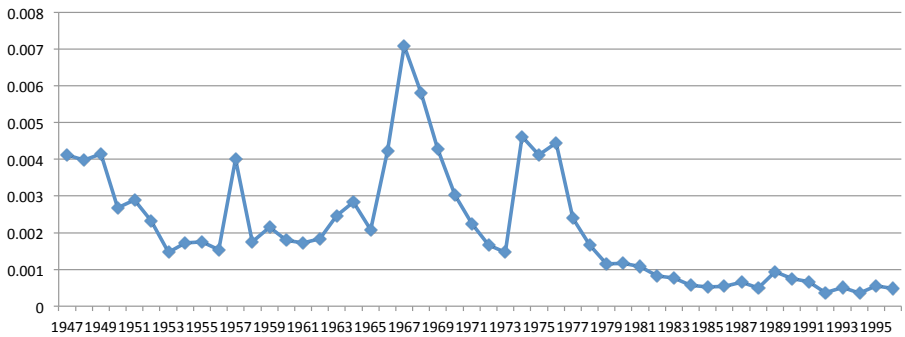

Fig. 2. The normalized frequency of words in the cluster of exploiting class and revolution in the diachronic corpus.

year(1967) with the highest frequency of these terms, and run a latent Dirichlet allocation (LDA) [2] on it to observe the topic distribution of these terms.

For each word in the produced cluster in Table 4, we observe its top 5 topics in LDA with the highest probability. We count the word number for each topic, and the relation between the frequencies of words with respect to the topics generated by LDA is shown in Figure 3. The terms in this cluster concentrates heavily on topic 17, 29, and 83. The words in those corresponding topics are shown in Table 5.

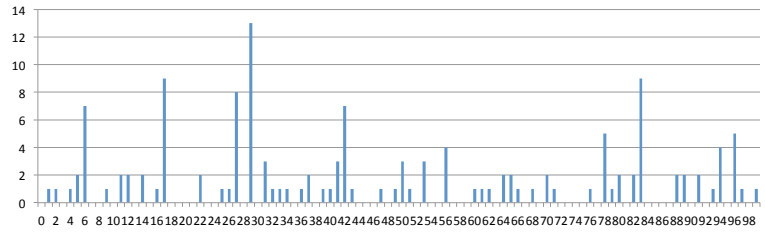

Fig. 3. Topic distribution of words in the cluster of exploiting class and revolution. y-axis indicates the frequency of words in a specific topic.

Table 5. Words in related LDA topics from *People's Daily*

Topic #17	党(Party) , 领导(lead) , 武装(armed) , 斗争(struggle) , 共产党(Communist party), 革命(revolution), 建立(establish) , 中国(China) , 政权(regime)
Topic #29	文化(culture) , 社会(society), 阶级(class), 思想(thoughts), 资产阶级(bourgeoisie), 生活(life), 制度(institution), 剥削阶级(exploiting class), 统治(rule)
Topic #83	印度(India) , 政府(government) , 武装(armed) , 地区(region), 反动(reactionary) , 农民(peasants), 军事(military), 挑衅(provoke), 巴基斯坦(Pakistan)

According to the deviation features and a hierarchical word clustering scheme[4], we could obtain clusters of hot topics. In this experiment, we found out that a generated cluster may correlate with several inter-related topics in LDA. Topic 17 talks about the armed partisanship revolution in China, topic 29 is about a political trend of anti-capitalism ideologies and topic 83 is about the military operations in other countries besides China. And the topic 6, 27 and 42, which the cluster also concentrated, are similar to these political topics. All these topics reflect the specific social background in that age.

4.4 Topic based diachronic analysis of social change

Michel et al. tried to investigate cultural trends quantitatively in a corpus of digitized texts containing about 4% of all books ever printed[6]. By tracking the

usage frequency of words picked carefully in different years, they highlight that the cultural change guides the concepts we discuss (hot topics). An example is shown in Figure 4, in which they plotted the median frequency in German over time for five lists of names and a collection of Nazi party members(547 names) to probe the impact of censorship on a person's cultural influence in Nazi Germany.

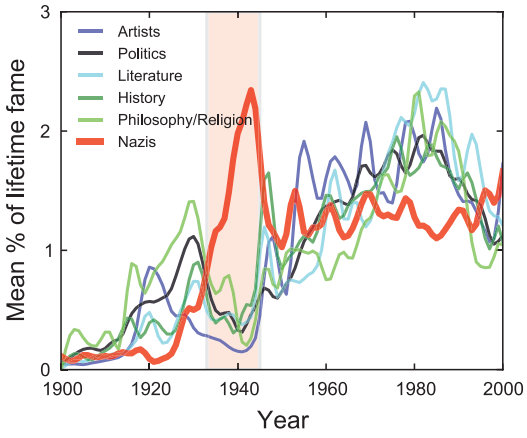

Fig. 4. Artists and writers in various disciplines were suppressed by the Nazi regime (red highlight). In contrast, the Nazis themselves (thick red line) exhibited a strong fame peak during the war years.[6].

Therefore, tracking the frequency of some certain words could detect the age of peak celebrity of some topics and study the human culture. While these special words were chose manually in Michel's work[6], by our deviation features and clustering method, the clusters with words belong to a same topic were generated automatically. With this method, the topic based diachronic analysis of social and culture change is more effective then.

We have shown an example above in Figure 2 (also the red line in Figure 5). As this cluster of exploiting class and revolution was active before the founding of the People's Republic, it had relatively high frequency. And the frequency of it peaked in 1967, which is the beginning of the Cultural Revolution. Since the Cultural Revolution eulogized revolutionary violence, and some claustrophobic restrictions was proposed, the political-related topic developed. After this period, with the reform and open policy carrying on, the revolution and military operations stepped down from the stage of history so that the frequency declined gradually.

Another example is shown in Figure 5, which is a cluster about literature and art. Its frequency was low during the war years but soared since the founding of PRC (1949), and retained high for nearly 20 years, but then underwent a rapid decay in 1967 reversely, dropping the bottom over 10 years. And returned to the average level before 1967. This is because the artistic creation was signifi-

cantly influenced by ideological factors in the Cultural Revolution (1967-1977) but reached its peak accompanying economic development.

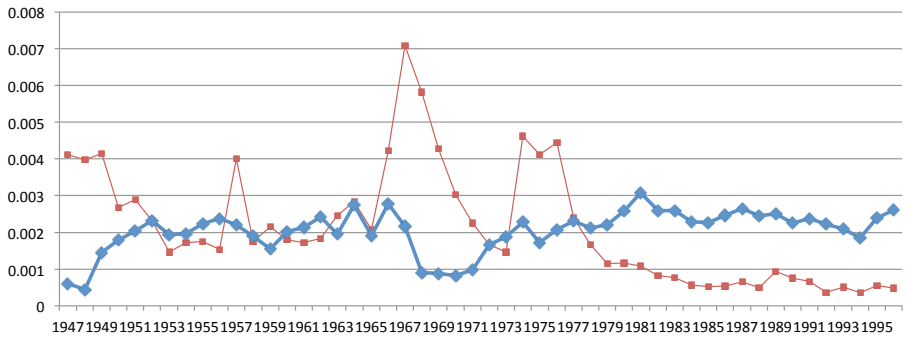

Fig. 5. The normalized frequency of the cluster "话剧(drama), 作品(words), 创作(creation), 戏剧(drama), 文学(literature), 演员(actor), 音乐(music), 连环画(comic strip), 绘画(drawing), 语言(language), 诗歌(poetry), 故事(story), 诗人(poet), 杂技(acrobatics), 现实主义(realism), 文艺(literature and art), 作家(writer), 电影(movie), 油画(painting), 歌剧(opera), 影片(film), 民歌(folk song), 小说(novel), 纪录片(documentary), 美术(art), 编排(arrange), 艺术(art), 鲁迅(Xun Lu)" (blue line) compare with the frequency of the cluster of exploiting class and revolution (red line).

This analysis illustrates that the frequency change of topics can reflect the culture change.

5 Conclusion and Future Work

This paper proposed deviation features of the diachronic word semantic distribution, which represents the stability of the word-meaning over time periods. The word semantic distribution can be learned via linear transformations of the word embedding results generated from the different time periods of text in the diachronic corpora. We demonstrated that the clustering result of words using these deviation vectors is correlated with time-specific topics. The frequency changes of these word clusters can indicate the social and culture changes in history.

Our future work includes exploiting this feature to other NLP tasks including diachronic topic mining or semantic change mining, and it can be also applied to the social linguistics and historical linguistics.

Acknowledgments. This work is supported by the NNSF of China (grant No. M1321005)

References

1. Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C.: A neural probabilistic language model. JMLR 3, 1137–1155 (2003)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. JMLR 3, 993–1022 (2003)
3. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. JMLR 12, 2493–2537 (2011)
4. He, S., Zou, X., Xiao, L., Hu, J.: Construction of diachronic ontologies from people's daily of fifty years. In: LREC (2014)
5. Kleinberg, J.M.: Hubs, authorities, and communities. ACM Computing Surveys 31(4es), 5 (1999)
6. Michel, J.B., Shen, Y.K., Aiden, A.P., Veres, A., Gray, M.K., Pickett, J.P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., et al.: Quantitative analysis of culture using millions of digitized books. science 331(6014), 176–182 (2011)
7. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
8. Mikolov, T., Le, Q.V., Sutskever, I.: Exploiting similarities among languages for machine translation. arXiv preprint arXiv:1309.4168 (2013)
9. Zhang, H.P., Yu, H.K., Xiong, D.Y., Liu, Q.: Hhmm-based chinese lexical analyzer ictclas. In: SIGHAN. pp. 184–187 (2003)