Word Semantic Variation Mining based on Pos-Word2vec and cloud-model

Maoyuan Zhang[1,a] ,Xiaohang Pan[2,b], Qiongyao Meng[3,c]

[1,2,3]School of Computer, Central China Normal University, Wuhan, 430079, China

[a]email: zhangmy@mail.ccnu.edu.cn, [b]email: panxiaohang_love@126.com

[c]email:1126186992@qq.com

Keywords: Word variation, Distributed word representations,Cloud model

Abstract. Words, as the basis of language, have relative stability.But it is also gradually developing. The language variation and change has been an important subfield in sociolinguistics, and has made remarkable achievements.But there is seldom research conducted from the aspect of the natural language processing.This paper applied the distributed word representations to map the word semantic,and used the cloud model to calculate the variation of the word.The experiment shows that our approach achieves better semantic information and useful results in semantic variation and change analysis.

Introduction

Words, as the basis of language, have relative stability. Language is developed continuously, as a component of language, the words have the development and change of the language at the same time.So,the meaning of words is relatively stable, but it is also gradually developing.The language variation and change has been an important subfield in sociolinguistics, and has made remarkable achievements[1]. However, the methods of language study adopted by sociolinguists are generally empirical investigation(which may contain qualitative or quantitative analysis),which are usually laborious and time consuming,and there is seldom such research conducted from the aspect of the natural language processing(NLP).In NLP,the methodology is a typical corpus based statistical method which relies on the context(lexical or syntactic) of the target words and gives their statistical trends in semantic or usage[2].As an important method of the natural language processing, corpus based statistical approach has gained extensive attentions. It was widely adopted in machine translation, dictionary construction, grammar research and so on.

In this paper, the distributed word representations are used to map the word semantic to the multidimensional space, and the similarity computation is used to find the similar word in the space.Then,we convert the word and its related words into three characteristic values of the cloud model.For the same word in different time periods, the degree of variation can be measured by calculating the overlap degree of cloud model.

The rest of this paper is organized as follows.Section 2 presents our method specifically. Section 3 describes the experiment.The last section concludes this paper and discusses the future work.

Word Semantic Variation Mining based on Method

In order to mining the word semantic change, we first need to know how to acquire the semantic knowledge of a word.this paper employs the cloud model to transfer one word to one concept by synonyms and relative words expansion for every word , and then adopts the concept synthesis method to obtain the concept quantification of the word.

Word Expansion based on Pos-Word2vec

Deep learning is the most rapid development in the field of machine learning in recent years.Strictly speaking, the deep learning is not a new machine learning methods, but based on the nickname of deep neural network method.With the breakthrough in the field of speech and image processing, deep learning has attracted more and more attention, and gradually applied to various tasks in NLP.Distributed representation is the application of deep learning in the field of NLP[3].Distributed representation is dense,low dimensional,and real-valued.Distributed word representations are called word embeddings. Each dimension of the embedding represents a latent feature of the word, hopefully capturing useful syntactic and semantic properties. A distributed representation is compact, in the sense that it can epresent an exponential number of clusters in the number of dimensions.Word embeddings are typically induced using neural language models, which use neural networks as the underlying predictive model. Historically, training and testing of neural language models has been slow, scaling as the size of the vocabulary for each model computation.However, many approaches have been proposed in recent years to eliminate that linear dependency on vocabulary size and allow scaling to very large training corpora[4].Word2vec is a deep learning model,which is mainly used to generate distributed word representation[5].

In the general training process, word2vec uses the continuous bag of word (CBOW)model and Skip-Gram model[6].The CBOW model consists of three layers: the input layer, the projection layer and the output layer.As shown in Figure 1, we can see that the CBOW model uses the context phrase to predict the probability of the occurrence of the word.

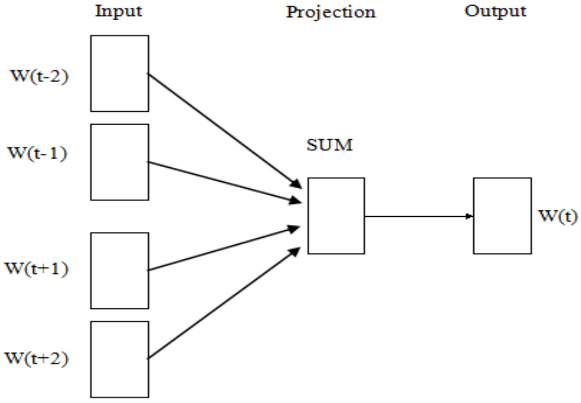

Fig.1. The CBOW model

The mapping process of the input layer to the projection layer is to add the word vector of all the words in the context.The model is simple in structure and easy to calculate, but it loses the feature information of the sentence.In order to obtain the sentence feature, we propose Pos-CBOW model.The difference between the model and the original word2vec is that the mapping of the input layer to the projection layer is no longer a simple accumulation, but combined with the lexical analysis tool Pos-tagger tools to extract the corresponding sentence features. Pos-tagger extract the core elements of the sentence, nouns, verbs, adjectives.In the input layer to the projection mapping layer, we give different weights for the different parts of speech.The calculation is described as Eq.(1).

$$X_w = \alpha w_{Noun} + \beta w_{Verb} + \gamma w_{Adj} \tag{1}$$

Where α, β, γ are different weight values.And $w_{Noun}, w_{Verb}, w_{Adj}$ are the sum of the vector of different parts.

We use hierarchical softmax to enhance the training speed of the model. In the context of neural network language models, it was first introduced by Morin and Bengio. The hierarchical softmax uses a binary tree representation of the output layer with the W words as its leaves and, for each node, explicitly represents the relative probabilities of its child nodes. These define a random walk that assigns probabilities to words.

For the vector v_{w1}, v_{w2} of W1 and W2, the semantic similarity between the two words is the cosine similarity of the two words, and the calculation is described as Eq.(2).

$$Similarity(w1.w2) = \frac{v_{w1} \cdot v_{w2}}{\|v_{w1}\| * \|v_{w2}\|} \tag{2}$$

For each word, we computed the twenty most similar entries using the cosine similarity.

Cloud Model Based Method: Transfer Word to Concept

Cloud model[7], unlike cloud computing, is a cognitive model in NLP.After a series of research on cognitive science, artificial intelligence and knowledge representation, Deyi Li proposes the theory of cloud model. From the perspective of natural language, it is more in line with the essence of things than the traditional mathematics. Fuzzy theory, probability theory, rough set theory and other basic theory to provide the foundation for the creation of the cloud model.It describes the uncertainties of linguistic concepts, especially the randomness and fuzziness, and implemented the uncertain transformation between linguistic concepts and quantitative values.A word and its relations constitute a concept. In order to express the concept, three quantifications of cloud model are introduced as follows.

(a)Concept Expectation: Concept Expectation is the mathematical expectation of the relative words belonging to a concept in the universal. It can be regarded as the most representative and typical sample of the qualitative concept.

(b) Concept Entropy: Concept Entropy represents the uncertainty measurement of a qualitative concept. It is determined by both the randomness and the fuzziness of the concept. As the measurement of randomness,it reflects the dispersing extent of the relative words. On the other hand, it is also the measurement of fuzziness, representing the scope of the universe that can be accepted by the concept.

(c) Hyper entropy: Hyper entropy is the uncertain degree of entropy.

Their equations are introduced as Eq.(3)

$$.\begin{cases} Ex = \frac{1}{l}\sum_{i=1}^{l} Xi \\ En = \frac{1}{l}\sqrt{\frac{\pi}{2}}\sum_{i=1}^{l}\left|Xi - Ex\right| \\ He = \sqrt{\left|S^2 - En^2\right|} \end{cases} \tag{3}$$

where Ex is the value of concept expectation, En is the value of concept entropy, He is the value of Hyper entropy, $S^2 = \frac{1}{l-1}\sum_{i=1}^{l}\left(Xi - Ex\right)^2$,l is the total number of relative words of center word, Xi is the relevance between the center word and its ith relative word . A concept can be expressed as C(Ex;En;He).

Considering the internal relations of Ex,En and He,the region Rs[Ex-En;Ex+En], Rs[Ex-2En;Ex+2En],Rs[Ex-3En;Ex+3En]contributes 68.28%,95.46%, 99.74% to the meaning of concept.When two words express or indicate approximately similar meaning, their quantification will be much near each other, so the rate of intersection over the whole two numerical regions reflects the similarity of two words. The higher the rate, the more similar the word pair is.Suppose vector C1(Ex1,En1,He1) stands for word w1, vector C2(Ex2,En2,He2) represents sentence w2, the region Rc [Ex -3En ;Ex+3En] represents concept C.Finally the score between C1 and c2 derived from Eq.(4) represents the variation degree of the word pair.

$$R_{SC}(c1, c2) = 1 - \frac{R_{C1} \cap R_{C2}}{R_{C1} \cup R_{C2}} \tag{4}$$

Experiment and Results

In this paper, we use the data collected from web news.We segment all of the text automatically by using the word segmentation software ICTCLAS, do preprocessing for all of the data, such as stopping word.Then We use the Pos-CBOW model to train the text. So the words will be represented as continuous vectors.In this representation, the vectors' weights are directly computed so as to maximize the probability of the context in which the word being modeled tends to appear.

This permits efficient representation of models trained on massive amounts of data in relatively small-sized vectors. We used the 200-dimensional vectors trained on the 100 million-word Sina News dataset .The model covers more than 180 thousands words and phrases, which is a considerable vocabulary size. For each entry, we computed the twenty most similar entries using the cosine similarity. We calculate the variation degree of all words,take out the top ten.Table 1 display the top 10 sensitive words.

Table 1. The top 10 sensitive words

Ranking	word	semantic change	Ranking	word	semantic change
1.	Lower Middles	0.960820	6.	Arguing	0.895674
2.	greeting	0.932811	7.	be lucky	0.892654
3.	Consume	0.910377	8.	miserly	0.891965
4.	Manufacturing	0.899459	9.	embarrassed	0.889548
5.	Intelligence	0.897254	10.	party	0.885626

Conclusions and Future Work

This paper studies word semantic variation and change mining from the aspect of computational lexical semantics.The preliminary experiments show that our appproach achieves a helpful result in words semantic variation and change analysis in both overall trends and word level characteristics.Our future work will focus on the following aspects.Firstly,using more refined algorithm to process the corpus and designing more elaborate model in word semantic mining.Secondly,many other social historical changes mining can be conducted based on the corpus.

Acknowledgement

In this paper,the research was sponsored by the Humanity and Social Science Youth Foundation of Ministry of Education of China (Project No. 15YJC870029) and National Social Science Foundation (Project No. 12&2D223).

References

[1] Walker J A. The Handbook of Language Variation and Change (review)[J]. Language, 2004, 80(3):591-594.

[2] Jurafsky D, Martin J H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition[J]. Prentice Hall, 2008, 26(4):638-641.

[3] Hinton, Geoffrey E. Learning distributed representations of concepts[C]. Proceedings of the eighth annual conference of the cognitive science society. 1986:1-12.

[4] Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning[J]. International Conference on Machine Learning, 2008:160-167.

[5] Mikolov T, Sutskever I, Chen K, et al. Distributed Representations of Words and Phrases and their Compositionality[J]. Advances in Neural Information Processing Systems, 2013, 26:3111-3119.

[6] BJ Turian, DD Et, Recherche Operationnelle.Word representations: A simple and general method for semisupervised learning.In ACL 178-1.

[7] Li D, Liu C, Gan W. A new cognitive model: Cloud model[J]. International Journal of Intelligent Systems, 2009, 24(3):357-375.