Research About Zipf's Law and Information Entropy on Chinese Text

Zhibo Liang

zbliang_sasee@buaa.edu.cn

Abstract. This report performs a research about semantic training on Chinese Text. A total of sixteen books are utilized to operate the training and testing. As a result, the performance of semantic training based on word2vec is acquired.

1 Introduction

Semantic training is a machine learning technique used to train artificial intelligence systems to understand the meaning and context of natural language. The main purpose of semantic training is to enable AI systems to better understand human language, rather than just recognizing words. This includes understanding metaphors, context, tone, and other semantic meanings. Semantic training plays a crucial role in many AI applications, such as dialogue systems, question-answering systems, text summary, etc. Enabling systems to better understand semantics helps improve the performance of these applications. The main challenges in semantic training include the complexity, diversity, and ambiguity of language. Accurately capturing semantics requires large amounts of training data and advanced language understanding models. In summary, semantic training is an important machine learning technique that helps advance natural language processing and the development of artificial intelligence.

2 Methodology

2.1 Word2vec

Word2vec is a neural network-based model that learns to represent words as numerical vectors, called word embeddings. The key idea behind word2vec is that words with similar meanings will have similar vector representations. There are two main approaches to training word2vec models:

Skip-gram model[1]: This model tries to predict the surrounding words given a target word. The intuition is that words appearing in similar contexts will have similar vector representations.

Continuous Bag-of-Words (CBOW) model[2]: This model tries to predict a target word given its surrounding context words. This essentially works in the opposite direction of the skip-gram model.

2.2 Cosine Similarity

Cosine similarity is a commonly used text similarity calculation method, mainly used in fields such as information retrieval and text classification. It measures the similarity between two vectors by calculating the cosine of the angle between them. It can be represented as follows[3]:

$$\frac{\boldsymbol{x}^\top \boldsymbol{y}}{||\boldsymbol{x}||||\boldsymbol{y}||} \in [-1, 1] \tag{1}$$

The range of cosine similarity values is [-1, 1]. A value closer to 1 indicates that the two vectors are more similar, 0 indicates that they are completely orthogonal, and -1 indicates that they are completely opposite.

3 Experiment Results

3.1 Training Result

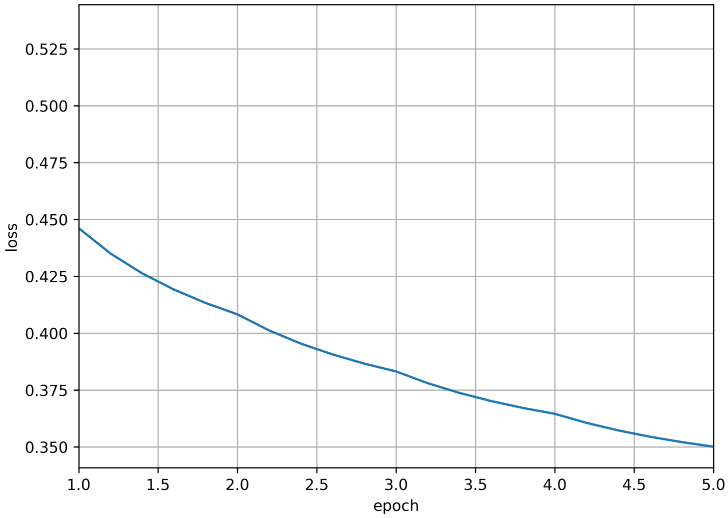

Fig. 1: The training result of word2vec.

The training process for word2vec involves feeding a large corpus of text data into the neural network. As the network is trained, it learns to encode semantic and syntactic relationships between words into the vector representations. The model used in this research is Skip-gram model and cosine similarity is utilized to conduct similarity assessment. Besides, loss function selected for this training is binary cross-entropy loss. The training result is shown as Fig. 1.

The testing result, based on cosine similarity retrieval, which selects the words with the highest cosine similarity, is shown in Fig. 2.

Input token	1st similar token	2nd similar token	3rd similar token
鹿	羚	麋	笃
少年	小伙子	十五六岁	矮小
野兽	猛兽	十多头	黄羊
兄弟	结义为	难共当	朋友

Fig. 2: The details of the testing result.

4 Conclusion

This report conducted a simple study on semantic training based on word2vec. The training and testing results show that this method can effectively perform semantic modeling and training. At the same time, it also proves to a certain extent that cosine similarity has a good ability to represent semantic similarity.

References

1. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” *Advances in neural information processing systems*, vol. 26, 2013.
2. T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *arXiv preprint arXiv:1301.3781*, 2013.
3. F. Rahutomo, T. Kitasuka, M. Aritsugi *et al.*, “Semantic cosine similarity,” in *The 7th international student conference on advanced science and technology ICAST*, vol. 4, no. 1. University of Seoul South Korea, 2012, p. 1.