Natural language inference by deep learning method

Saihan Li*, *Zhijie* Hu and *Rong* Cao

School of Information Engineering, Xi'an Eurasia University, Xi'an, China

Abstract. Natural Language inference refers to the problem of determining the relationships between a premise and a hypothesis, it is an emerging area of natural language processing. The paper uses deep learning methods to complete natural language inference task. The dataset includes 3GPP dataset and SNLI dataset. Gensim library is used to get the word embeddings, there are 2 methods which are word2vec and doc2vec to map the sentence to array. 2 deep learning models DNNClassifier and Attention are implemented separately to classify the relationship between the proposals from the telecommunication area dataset. The highest accuracy of the experiment is 88% and we found that the quality of the dataset decided the upper bound of the accuracy.

Keywords: Natural language inference, Deep learning, Word2vec.

1 Introduction

Deep learning which has multiple layers processing structure has blossomed in different areas such as computer vision, speech recognition and bioinformatics. In the recent years, deep learning has shown its outstanding accuracy and efficiency in natural language processing area [1]. Natural language processing is an area of using technical methods to deal with human natural language. The task of natural language processing includes tokenization, part-of-speech tagging, spelling error identification and so forth [2].

Natural language processing method change human language to numeral vectors for machine to calculate. With these word embeddings, researchers can do different task such as sentiment analysis, machine translation and natural language inference. Natural language inference (NLI) — characterizing and using the relations in computational systems is essential in tasks ranging from information retrieval, semantic parsing to commonsense reasoning [3].

Natural language inference is the new increasing area of natural language processing, which infer the relationship of 2 sentence, one is premise and the other one is hypothesis. The relationship between premise and hypothesis can be entail, contradictory and neutral[3,4]. For example, there are 4 sentences below, the task is trying to infer the relationship between the first sentence and the next 3 sentences.

* Corresponding author: lisaihan@eurasia.edu

If you help the needy, God will get reward.
Giving money to the poor has good consequences.
Giving money to the poor has unfortunate consequences.
It's sunny outside

The premise "If you help the needy, God will get reward.", the hypothesis "Giving money to the poor has good consequences.", the premise will entail the hypothesis. If the hypothesis changes to "Giving money to the poor has unfortunate consequences.", then they have contradictory statement. If the hypothesis is "it is sunny outside", the relationship between them is neutral.

2 Dataset

The dataset of this paper is combined by 2 parts which are 3GPP dataset and SNLI dataset

2.1 3GPP dataset

In this project, we want to see the relationship of 2 sentences from telecommunication area. The original source of the dataset is 3GPP meeting proposals, The sample of the dataset is as below:

Premise:

"R2-1802039: The SIA should be smaller than or equal to the TA and the TAI (broadcasted in MSI) +SIAID (SIA index in a TA) can uniquely identify an area."

Hypothesis:

R2-1802315: Each cell belongs to only one System Information Area"

2.2 SNLI dataset

The SNLI (Sandford Natural Language Inference) corpus (version 1.0) is a collection of 570k human-written English sentence pairs manually labelled for classification with the label "entailment, contradiction, and neutral", supporting the task of natural language inference. It includes the original 2 sentences, the structure of the 2 sentences and the label of the relationship. For most paper in natural language inference area, they use SNLI dataset as their training set for research.

3 Methods and models

3.1 Word embeddings

Words are map into vectors using word embedding models, the vectors are called word embeddings, they capture both semantic and syntactic information of words. Word embeddings can be used to calculate word similarities which can be used in many tasks such as information retrieval [5].

In this paper, 2 methods which are doc2vec and word2vec are used to accomplish word embeddings. In Gensim library, doc2vec model changes text document to numeric representations. Each sentence is one vector, the dimension of the vector can be determined by programmer. word2ve model maps each word to a vector. Therefore, the sentence length can be varied depending on the number of words in each sentence.

3.2 Deep learning models

There are 2 deep learning models implemented in this paper on the platform of Tensorflow which is an open-source software library for high performance numerical computation [7]. It is widely used in deep learning and machine learning tasks. The first one is using doc2vec model to change sentence to vectors and then use these vectors as input to a DNNclassifier which is used to predict the relationship of 2 sentences.

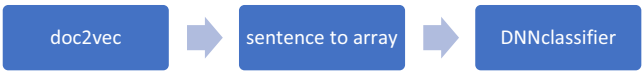

Fig. 1. Model_1.

In model_2, first we use word2vec to map each work to a vector, so the 2 sentences are different length. Then we chose decomposable attention model which was first introduced in [8] to classify. The model has 3 steps "attend, compare and aggregate". Each step is a 2 layers neural network.

In the first step "Attend", for 2 sentence s1 and s2, which is already replaced by word embeddings, soft-align the elements in 2 sentence vectors. s1=(a1, a2,…,a l_a), s2=(b1,b2,…,bl_b). First, a factor e is calculated as the function below:

$$e_{ij} = F(a_i)^T F(b_j)$$

F function here is a 2layer feedforward neural network with Relu activation. And then the attention weight is calculated for each word.

$$\beta_i = \sum_{j=1}^{l_b} \frac{\exp(e_{ij})}{\sum_{k=1}^{l_b} exp(e_{ik})} b_j$$

$$\alpha_j = \sum_{i=1}^{l_a} \frac{\exp(e_{ij})}{\sum_{k=1}^{l_a} \exp(e_{kj})} a_i$$

Here, β_i is the word in s2 which is soft-aligned to a_i in s1, and α_j is the subphrase is s1 which is soft-aligned to b_j in s2.

In the "compare" step, after we got the soft-aligned vector, using a function G which is also a 2 layer fully connected neural network. The input for G is the concatenation of the original vector and the corresponding attention weight.

$$c_i = con\langle a_i, \beta_i \rangle$$

$$c_j = con\langle b_j, \alpha_j \rangle$$

Con⟨,⟩ function here is concatenation of 2 elements. Then

$$v_{1i} = G(c_i) \quad \forall i \in [1, \dots, la]$$

$$v_{2j} = G(c_j) \quad \forall j \in [1, \dots, l_b]$$

The last step is "aggregate", it first aggregates over each column sum as below:

$$v_1 = \sum_{i=1}^{la} v_{1,i} \qquad v_2 = \sum_{j=1}^{lb} v_{2,j}$$

Then the result was fed into a classifier which is a 2 layer feedforward neural network with relu as first layer and use softmax cross-entropy as its loss function. It is a 2-layer feedforward neural network followed by a linear layer. [6] The model structure is

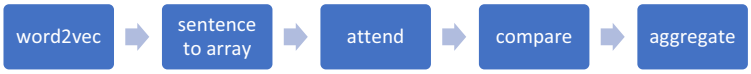

Fig. 2. Model_2.

4 Experiments

We extract premise, hypothesis and the label from every SNLI dataset record at the beginning. In training the language model, since the 3GPP training dataset is so limited, we combine the dataset of SNLI (Sandford Natural Language Inference) dataset and 3GPP dataset to train the language model. Then we exploit the model to transfer every sentence in the 3GPP training set to 300-dimension vectors. The vectors are input to DNNClassifier which is a classifier for TensorFlow DNN models. The test accuracy is 88%

When testing, we found the highest test accuracy is 88%, and can't be improved by increasing more epochs, it's because over 80% training data labels are Neutral so it's not balanced for 3 classes, and there are 12% test data labels are not 'Neutral', and they can't be predicted to a right value, which means all the Neutral labels test case are predicted right.

To overcome this drawback, oversampling method is taken. Because the training dataset is not balanced for 3 classes, we multiple the minority classes by 6 times to make 3 classes balanced. The test accuracy is 57.3%. The reason is that some 'Neutral' label test cases are predicted wrong since there are more training sample with labels 'Entailment'. That's why the accuracy is even less than before oversampling.

In the second experiment, the same dataset is adopted to train the word2vec model, the model will infer every word to a 300-dimension vector, so the length of the sentence will determine the dimension of the word embedding matrix, but the column number for all the sentence should be same which is the fixed dimension of the word embedding. Then these vectors are input to decomposable attention model. The accuracy is also 88% which is the same as the first experiment. The reason is the same, because the test training set is so limited.

One drawback of word2vec model is that if there is a new word which is not in the language model training set, there will be an error, so the training set must include all the words. One solution to the OOV(out of vocabulary) problem would be hash the new word to a vector [6].

5 Conclusion

This paper implemented 2 deep learning methods to explore Natural Language Inference in the area of telecommunication. From the result we can see the data decides the upper bound of the result, and the 2 methods are performing well in predicting the majority class but not in the minority classes in classification.

The research was sponsored by the Special scientific research plan of Education Department of Shaanxi Provincial Government (Project No. 21JK0817) and Social science fund of Shaanxi Province (Project No.2019Q019).

Reference

1. Young, Tom et al. “Recent Trends in Deep Learning Based Natural Language Processing [Review Article].” IEEE Computational Intelligence Magazine 13 (2018): 55-75.
2. Prakash M Nadkarni, Lucila Ohno-Machado, Wendy W Chapman; Natural language processing: an introduction, Journal of the American Medical Informatics Association, Volume 18, Issue 5, 1 September 2011, Pages 544–551
3. Cleo Condoravdi, Dick Crouch, Valeria de Paiva, Reinhard Stolle, and Daniel G. Bobrow. Entailment, intensionality and text understanding. In Proc. of the HLT-NAACL 2003 Workshop on Text Meaning.
4. R. Bora-Kathariya and Y. Haribhakta, "Natural Language Inference as an Evaluation Measure for Abstractive Summarization," 2018 4th International Conference for Convergence in Technology (I2CT), 2018, pp. 1-4, doi: 10.1109/I2CT42659.2018.9057819.
5. Samuel R. Bowman, Gabor Angeli,Christopher Potts, and Christopher D. Manning. 2015.A large annotated corpus for learning natural language inference. In Proceedings of EMNLP
6. Yang Liu , Zhiyuan Liu , Tat-Seng Chua , Maosong Sun. Topical Word Embeddings. Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence
7. Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2016. TensorFlow: a system for large-scale machine learning. In Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation (OSDI'16). USENIX Association, Berkeley, CA, USA, 265-283.
8. Ankur Parikh, Oscar Täckström, Dipanjan Das, Jakob Uszkoreit . A Decomposable Attention Model for Natural Language Inference. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing ,2016. Association for Computational Linguistics, 2249-2255.