An Improved LSTM Structure for Natural Language Processing

Lirong Yao
Qingdao No.2 Middle School
Qingdao, China
lryao_cn@163.com

Yazhuo Guan
Intelligence and Big Data Research Center
Global Wisdom Inc.
Beijing, China.
guanyazhuo@langlib.com

***Abstract*—Natural language processing technology is widely used in artificial intelligence fields such as machine translation, human-computer interaction and speech recognition. Natural language processing is a daunting task due to the variability, ambiguity and context-dependent interpretation of human language. The current deep learning technology has made great progress in NLP technology. However, many NLP systems still have practical problems, such as high training complexity, computational difficulties in large-scale content scenarios, high retrieval complexity and lack of probabilistic significance. This paper proposes an improved NLP method based on long short-term memory (LSTM) structure, whose parameters are randomly discarded when they are passed backwards in the recursive projection layer. Compared with baseline and other LSTM, the improved method has better F1 score results on the Wall Street Journal dataset, including the word2vec word vector and the one-hot word vector, which indicates that our method is more suitable for NLP in limited computing resources and high amount of data.**

Keywords—NLP, LSTM, machine translation, deep learning, recursive projection layer

I. Introduction

Computational linguistics, also known as natural language processing (NLP), is a subfield of computer science involving the use of computational techniques to learn, understand, and produce human language content. Computational language systems can be used for a variety of purposes: goals can help people communicate, such as machine translation; help with human-computer interactions, such as with conversational agents; by analyzing and learning many human languages, making humans and Machine benefit from content that is now available online. In the first decades of computational linguistics, scientists attempted to write vocabulary and rules for human languages for computers. This has proven to be a daunting task due to the variability, ambiguity and context-dependent interpretation of human language [1].

In terms of research content, natural language processing includes grammar analysis, semantic analysis, and text comprehension, which is shown in Fig.1. The most well-known applications of natural language processing include Google's knowledge map, IBM Watson's natural language question and answer, and Apple Siri's dialogue system. NLP involves data mining, machine learning, knowledge acquisition, knowledge engineering, artificial intelligence research, and linguistic research related to language processing. Advances in natural language processing technology have played an important role in machine translation applications. Machine translation refers to the translation of a document from one language to another by a computer, and a one-to-one correspondence between grammatical structure and vocabulary identification. Accurate machine translation can greatly improve the efficiency of human social communication and understanding.

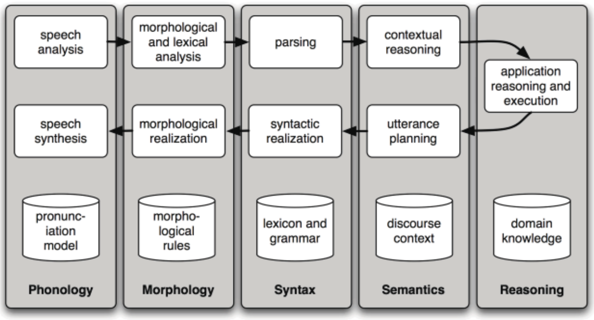

Figure 1. The structure of natural language analysis

N-gram based technology dominates modern natural language processing (NLP) and its applications. Typically, they are used as features to represent a vector space model, and then a standard classification algorithm is applied to the model. The value of these features is a frequency of n grams and may be weighted in some way. A traditional n-gram is a sequence of elements that appear in text. These elements can be words, characters, POS tags, 1 or any other element as they appear one after another in the text. The usual convention is that 'n' in an n-gram corresponds to the number of elements in the sequence [2]. Natural language analysis is broken down into various levels, such as phonological, morphological, syntactic, semantic, pragmatic and discourse analysis [3].

- Phonology

Phonology is an analysis of spoken language. Therefore, it involves speech recognition and generation. The core task of the speech recognition and generation system is to take the acoustic waveform as input and generate a string of words. Phonology is part of natural language analysis and it deals with it. The field of computational linguistics that deals with speech

978-1-5386-5514-6/18/$31.00 ©2018 IEEE

analysis is computational phonology.

- Morphology

This is the most basic stage of NLP. It handles word formation. At this stage, a single word is analyzed according to a component called a "morpheme." In addition, non-words such as punctuation are separated from words. Morphemes are the basic grammar building blocks that make up a word.

- Syntax

Grammar refers to the study of the formal relationship between sentence words. At this stage, the validity of the sentence according to the grammar rules is checked. To perform parsing, you need grammar and analytical knowledge. Grammar is the formal specification of the rules allowed in the language. Parsing is a method of analyzing sentences according to their grammar to determine their structure. The most common grammar used for natural language parsing is the contextless grammar (CFG), also known as the stage structure grammar and the explicit sentence grammar. These grammars are described in detail in separate actions.

- Semantics

In linguistics, semantic analysis is the process of linking syntactic structures from the levels of phrases, clauses, sentences, and paragraphs to the entire level of writing, as well as language-independent meaning. It also involves deleting features that are specific to a language and cultural context, if such a project is possible. The elements of idioms and figurative language, that is, culture, are often transformed into relatively invariant meanings in semantic analysis. Semantics, although related to pragmatics, is unique in that the former deals with word or sentence choices in any given context, while pragmatics considers unique or specific meanings derived from context or intonation. Reaffirmed in different terms, semantics is about the meaning of universal coding, pragmatics is the meaning of coding in words, and then interpreted by the audience.

Natural Language Processing (NLP) benefits from the renaissance of Deep Neural Networks (DNN) because of their high performance and low demand for engineering features. There are two main DNN architectures: Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN). Gating mechanisms have been developed to mitigate some of the limitations of basic RNN, resulting in two main types of RNN: long short-term memory (LSTM) and gated recurrence units (GRU). In general, CNN is a hierarchical and RNN sequential architecture. How should we choose them to handle the language? Based on the characteristics of CNN and RNN, it is tempting to select CNN as a classification task such as sentiment classification, because emotions are usually determined by some key phrases; and RNN is selected for sequence modeling tasks such as language. Modeling because it requires flexible modeling of context dependencies. But the current NLP literature does not support such a clear conclusion. For example, RNN performs well in document-level sentiment classification. It has recently been shown that closed CNN performs better than LSTM in language modeling tasks, although LSTM has long been considered more suitable. In summary, there is no consensus on DNN selection for any NLP issue [4].

Deep learning architectures and algorithms have made remarkable advances in areas such as computer vision and pattern recognition. Following this trend, recent NLP research is now increasingly focused on using new deep learning methods. For decades, machine learning methods for NLP problems have been based on shallow models such as SVM and logistic regression trained on very high dimensions and sparse features. In the past few years, neural networks based on dense vector representation have produced excellent results on various NLP tasks. This trend is triggered by the success of embedded words and deep learning methods. Deep learning enables multi-level automatic feature representation learning. In contrast, traditional machine learning-based NLP systems rely heavily on hand-crafted functionality. This handcrafted feature is very time consuming and often incomplete.

It is proved that a simple deep learning framework outperforms most of the most advanced methods in several NLP tasks, such as named entity recognition (NER), semantic role tag (SRL), and POS tagging. Since then, many sophisticated deep learning-based algorithms have been proposed to solve difficult NLP tasks. Statistical NLP has become the main choice for modeling complex natural language tasks. However, at the beginning, it is often accustomed to suffering from the notorious dimensional curse when learning the joint probability function of a language model. This leads to the motivation to learn the distributed representation of words that exist in low dimensional space [5].

However, the current NLP mission still has some shortcomings, although the current deep learning technology has made great progress in NLP technology. However, the current principal NLP system still has practical problems such as high training complexity, difficulty in calculation in massive content scenarios, high retrieval complexity, and lack of probability meaning. This paper proposes an improved NLP method based on long short-term memory (LSTM) structure. Compared with other NLP scheme, our method is more suitable for NLP in the case of training computing resources and complex retrieval, and has good performance in the case of high input data volume.

The structure of this article is arranged as follows: Firstly, several main deep learning-based NLP algorithms are given. Then an improved NLP method based on long short-term memory (LSTM) structure is given, and the method is given on the standard NLP data set. The performance indicators and test results, and finally the article summary.

II. Natural Language Processing based on Deep Learning

Since 2010, deep learning has developed rapidly, covering areas such as image recognition, speech recognition, natural language understanding, OCR, etc. Almost all traditional pattern recognition methods can be replaced by deep learning methods. Deep learning has evolved from traditional DNN to CNN, RNN and LSTM. These methods have achieved many good results in voice, image and text. Natural Language Processing is one of the important areas that CNN and LSTM are widely adopted.

A. Language model based on deep learning

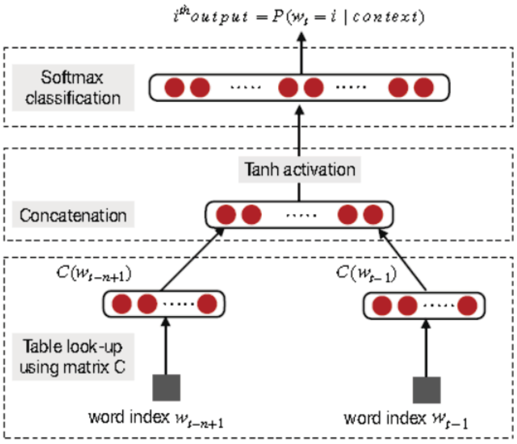

Figure 2. Bengio's Neural Language Model

In 2003, Bengio proposed a neural language model that can learn the distributed representation of words (shown in Fig.2). The authors believe that these words represent an exponential number of semantically adjacent sentences once they are compiled into a sentence representation using the joint probability of the word sequence. This, in turn, facilitates promotion, because if you have seen a sequence of words with similar words (about nearby word representations), invisible sentences can now collect higher confidence.

Mikolov proposed the CBOW and skip-gram models. CBOW calculates the conditional probability of the target word, given that the context word surrounding it is on a window of size k. On the other hand, the skip-gram model is exactly the opposite of the CBOW model by predicting the surrounding context words for a given central target word. It is assumed that the context words are symmetrically positioned with the target word within a distance equal to the size of the window in both directions. In an unsupervised setup, the word embedding dimension is determined by the accuracy of the prediction. As the embedding dimension increases, the accuracy of the prediction increases until it converges at a certain point, which is considered the best embedding dimension because it is the shortest without affecting accuracy [6].

B. CNN-based Natural Language Processing

CNN is a layered neural network whose convolutional layer alternates with the sub-sampling layer, reminiscent of simple and complex cells in the primary visual cortex. The CNN family differs in the implementation of convolution and sub-sampling layers and in the way the network is trained. The convolutional layer is parameterized by the size and number of maps, kernel size, skip factor, and join table. Each layer has M mappings of the same size. The kernel shifts over the active area of the input image. The skip factor defines the number of pixels that the filter/core skips in the x and y directions between subsequent convolutions [7].

In Natural Language Processing, one shallow neural network was proposed on six data sets, especially the Stanford Emotional Tree Library (SST). It consists of a convolutional layer followed by a maximum pool level. The final classifier uses a fully connected layer with drop-out. In paper [8], it shows that performance improves with increased depth, using up to 29 convolutional layers, which is shown in Fig.3. The model begins with a lookup table that creates a vector representation of each character and converts the input sentence into a two-dimensional tensor of size f x s, where f is the dimension of the embedded space and s is the number of characters. Enter the text. In this work, the input is a fixed-size padding text: f can be thought of as the RGB dimension of an image equal to 1 x s size.

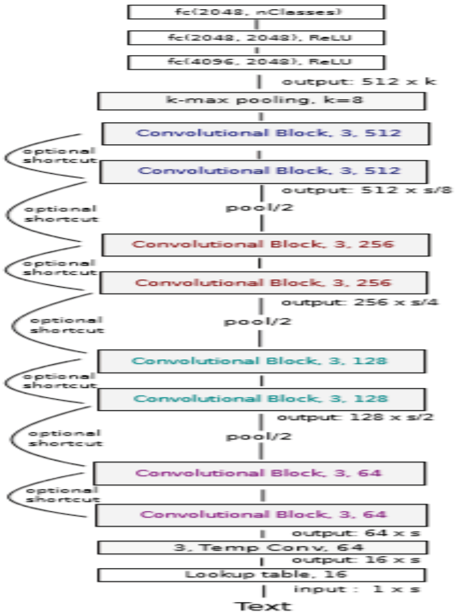

Figure 3. Global architecture with convolutional blocks.

C. LSTM-based Natural Language Processing

Long short-term memory is a recurring deep learning architecture that uses memory cell vectors and a set of element multiplication gates to control the way information is stored, forgotten, and utilized in the network. In the various connection designs of the LSTM unit, the architecture used in paper [9] is shown in Figure 4 and is defined by the following equations:

$$i_t = \sigma(\mathrm{W}_{wi}\, w_t + \mathrm{W}_{hi}\, h_{t-1}) \quad (1)$$

$$f_t = \sigma(\mathrm{W}_{wf}\, w_t + \mathrm{W}_{hf}\, h_{t-1}) \quad (2)$$

$$o_t = \sigma(\mathrm{W}_{wo}\, w_t + \mathrm{W}_{ho}\, h_{t-1}) \quad (3)$$

$$\hat{c}_t = \tanh(\mathrm{W}_{wc}\, w_t + \mathrm{W}_{hc}\, h_{t-1})t \quad (4)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \hat{c}_t \quad (5)$$

$$h_t = o_t \odot \tanh(c_t) \quad (6)$$

Where σ is the sigmoid function, i_t , f_t , o_t are input, forget, and output gates respectively, $\hat{c}_t$ and c_t are proposed cell value and true cell value at time t . Note that each of these vectors has a dimension equal to the hidden layer h .

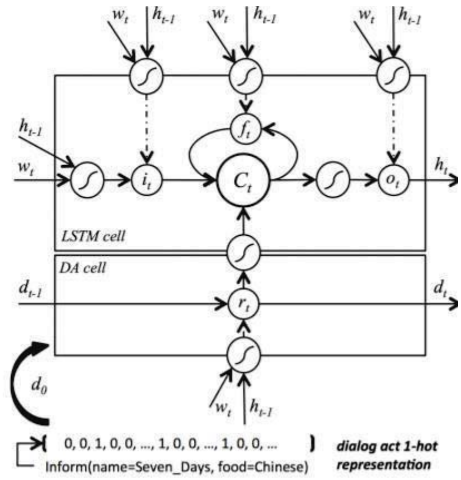

Figure 4. Semantic Controlled LSTM cell.

III. Improved NLP Method based on Long Short-Term Memory

In order to overcome the above problems of high training complexity, computational difficulty and high retrieval complexity, and to realize the requirements of training computing resources and complex retrieval, this paper proposes an improved NLP method based on long short-term memory (LSTM) structure. The composition is as shown in Fig. 5.

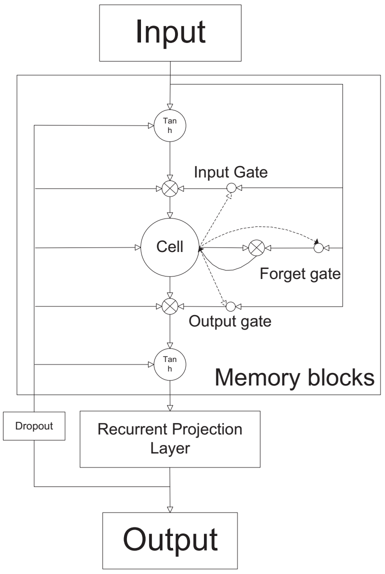

Figure 5. Improved LSTM for Natural Language Processing.

Recurrent Neural Network (RNN) is a neural network with a directed loop to express the dynamic time characteristics of the network. It is widely used in handwriting recognition and language modeling. Natural language texts have complex correlations on different time scales, so the cyclical neural network has a loop-connecting function that is more suitable for processing such complex time series data than deep neural networks. However, the gradient descent method used in RNN has the vanishing gradient problem, that is, in the process of adjusting the weight of the network, as the number of network layers increases, the gradient dissipates layer by layer, which makes it less and less effective for weight adjustment.

As a type of RNN, the LSTM model is more suitable than a cyclic neural network to process and predict long-term sequences with event lags and uncertain time [10]. LSTM adds a Recurrent Projection Layer to the RNN model to solve the gradient dissipation problem. The method proposed in this paper further adds the dropout module, which is used to randomly discard some parameter values when the parameters are passed backwards in the Recurrent Projection Layer, which reduces the training resource requirements and meets the fast detection requirements of real-time complex retrieval.

IV. Experimental Results and Analysis

To evaluate the performance of the above methods, we evaluated and tested them on our experimental computing environment. The specific configuration of the experiment is: three NVIDIA Tesla K40 GPU, the operating system is ubuntu17.10, all the implementations were one using TensorFlow.

Wall Street Journal dataset [11] is one of the machine-readable texts datasets over the last three years to generate meaningful statistical benchmark language models (including two-letters and triples). By changing the type of the selected language model, it is possible to evaluate the impact of the recognition performance of the variable confusion of the same text material. The availability of this text provides valuable resources to enable the development and evaluation of new language models and language models that are suitable for other tasks.

The splits of 70% sections for training, 10% sections for development and 20% sections for testing were adopted to imply our word level tagging task, where word2vec word vector (50-dimension) and one-hot word vector (512-dimension) was used to have contextual representations through our improved LSTM model (1024 units/256 projection), and the improved LSTM is used to encode all words before the prediction word, and the output vector was to predict the probability distribution of this word's occurrence through soft-max function.

Mean F1 score [12] shown in eq.7 was utilized to compare the performance of classic LSTM and our improved LSTM NLP method.

$$F_1 = 2\frac{precision \cdot \mathrm{recall}}{precision + \mathrm{recall}} \quad (7)$$

TABLE I. F1 SCORES WITH DIFFERENT NLP METHODS WITH WORD2VEC FEATURES

Compared Methods	$F_1 \pm \text{std}$
Baseline (DNN)	85.76%
Classic LSTM	87.41%
Classic BiLSTM	88.72%
Improved LSTM	88.93%

TABLE II. F1 SCORES WITH DIFFERENT NLP METHODS WITH ONE-HOT FEATURES

Compared Methods	$F_1 \pm \text{std}$
Baseline (DNN)	86.51%
Classic LSTM	88.96%
Classic BiLSTM	89.35 %
Improved LSTM	90.17%

In Table.I, the F1 scores have slowly rising from baseline to the improved LSTM method, which shows that the latter had 3.7%, 1.7% and 0.2% increase compared to other methods. In Table.II, with one-hot word vector, the F1 scores have similar rising speed from baseline to the new method, which shows that the latter had 4.2%, 1.3% and 0.9% increase compared to other methods.

V. CONCLUSION

This paper constructs an improved NLP method based on long short-term memory (LSTM) structure, in which randomly discard some parameter values when the parameters are passed backwards in the Recurrent Projection Layer. Compared with baseline and other LSTM, the improved method has better F1 score results on Wall Street Journal dataset both word2vec word vector and one-hot word vector, which implied our method is more suitable for NLP in the limited computing resources and high input data volume.

REFERENCES

[1] Schberg J, Manning C D. Advances in natural language processing[J]. Science, 2015, 349(6245): 261-266.

[2] Sidorov G, Velasquez F, Stamatatos E, et al. Syntactic n-grams as machine learning features for natural language processing[J]. Expert Systems with Applications, 2014, 41(3): 853-860.

[3] http://magizbox.com/training/natural_language_processing/site/tasks.html

[4] Yin W, Kann K, Yu M, et al. Comparative study of cnn and rnn for natural language processing[J]. arXiv preprint arXiv:1702.01923, 2017.

[5] Goldberg Y. A primer on neural network models for natural language processing[J]. Journal of Artificial Intelligence Research, 2016, 57: 345-420.

[6] Young T, Hazarika D, Poria S, et al. Recent trends in deep learning based natural language processing[J]. arXiv preprint arXiv:1708.02709, 2017.

[7] Ciresan D C, Meier U, Masci J, et al. Flexible, high performance convolutional neural networks for image classification[C]. IJCAI Proceedings-International Joint Conference on Artificial Intelligence. 2011, 22(1): 1237

[8] Conneau A, Schwenk H, Barrault L, et al. Very deep convolutional networks for natural language processing[J]. arXiv preprint, 2016.

[9] Wen T H, Gasic M, Mrksic N, et al. Semantically conditioned lstm-based natural language generation for spoken dialogue systems[J]. arXiv preprint arXiv:1508.01745, 2015.

[10] Zhang X, Zhao J, LeCun Y. Character-level convolutional networks for text classification[C]. Advances in neural information processing systems. 2015: 649-657.

[11] Paul D B, Baker J M. The design for the Wall Street Journal-based CSR corpus[C]. Proceedings of the workshop on Speech and Natural Language. Association for Computational Linguistics, 1992: 357-362.

[12] Yang Y, Liu X. A re-examination of text categorization methods[C]. Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1999: 42-49.