Advances in Intelligent Systems Research, volume 147

International Conference on Network, Communication, Computer Engineering (NCCE 2018)

Bidirectional LSTM-CRF Model and POS for Article Title Summarization

Xiaofeng Cai [1, a)], Zhifeng Hao [1,2]

[1] *Faculty of Computer Science, Guangdong University of Technology, Guangzhou 510006, China.*

[2] *Foshan University, Foshan 528000, China.*

[a)] Corresponding author: ms.xiaofengcai@hotmail.com

Abstract. In this paper, we propose a method based on Bidirectional LSTM-CRF for article title summarization. And for the summary generated by the model is not compliant with the grammar rules problem, we use POS (Part of Speech) to revise the generated summary. In order to verity our method, we conducted an experiment with the article title of WeChat public number. The results show that our method is effective, and POS can make results consistent with grammar rules and read fluently.

Key words: Bidirectional LSTM-CRF; title summarization; POS; word2vec.

INTRODUCTION

With the popularity of self-media age, More and more web articles have been published. Due to the diversity and freedom of online articles, the titles of online articles are also varied. Many online article titles are too long or there is redundant information. This will result in the reader not easily understanding the article title and reducing the reader's interest in the article. Therefore, it is of great significance to study article title simplification.

Article title simplification is a subtask of text summarization. And text summarization is the task of generating a short summary consisting of a few sentences that captures the main ideas of an article or a passage. In recent past, deep-learing based models that map an input sequence into another output sequence, called Bidirectional LSTM-CRF, have been successful in many problems such as sequence tagging [1] and named entity recognition [2].

In this paper, our goal is to simplify the title of article. To achieve this goal, we propose a method based on Bidirectional LSTM-CRF and POS. in our method, firstly, we use BIO to tag the text after word segmentation work.; secondly, we use Bidirectional LSTM-CRF to learn initial summarization;finaly, POS is used to revise the results that disobedience grammar rules.

Proposed Method

In this section, first, the WeChat public number article we collected was labeled using IOB scheme; second, we describe the Bidirectional LSTM-CRF model used to train the summarization generation model; finally, POS is used to revise the generated summarization.

Data Labeling

We collected 10000 article title form WeChat public number. The tagging scheme is the IOB scheme originally put forward by Ramshaw and Marcus (1995). Words tagged with O are outsize of title summarization and the I tag is used for words inside a title summarization. The first word of title summarization will be tagged B Here is an example sentence:

Fans in-depth Daisy's crew pictures are exposure, casually dressed, very kind!

Bidirectional LSTM-CRF Model

Long Short-Term Memory (LSTM) is a special type of Recurrent Neural Networks (RNN) that can learn long-term dependencies. Compared to the traditional recurrent neural network, there are two major changes. The first is the introduction of memory cells; the second is the mechanism of adding the gate. In LSTM, a sketch of a unit is shown in Figure 2. Formally, the update of each LSTM component can be formalized as

$$\begin{cases} i^t = \delta\left(W_i x^{(t)} + U_i h^{(t-1)} + b_i\right) \\ f^t = \delta\left(W_f x^{(t)} + U_f h^{(t-1)} + b_f\right) \\ o^t = \delta\left(W_o x^{(t)} + U_o h^{(t-1)} + b_o\right) \\ \tilde{c}^{(t)} = \tanh(W_c x^{(t)} + U_c h^{(t-1)} + b_c) \\ c^{(t)} = f^{(t)} \odot c^{(t)} + i^{(t)} \odot \tilde{c}^{(t)} \\ h^{(t)} = o^{(t)} \odot \tanh\left(c^{(t)}\right) \end{cases} \tag{1}$$

Where σ is the logistic sigmoid function Operator ⊙ is the pointwise multiplication of two vectors i^t, f^t, o^t, $c^{(t)}$ are the input gate, forget gate, output gate, and memory cell activation vector at time-step t respectively, all of them have the same size as the hidden vector $h^{(t)} \in R^H$. $W_i, W_f, W_o, W_c \in R^{Hxd}$ and $U, U_f, U_o, U_c \in R^{HxH}$ are trainable parameters. Here, H and dare the dimensionality of hidden layer and input respectively.

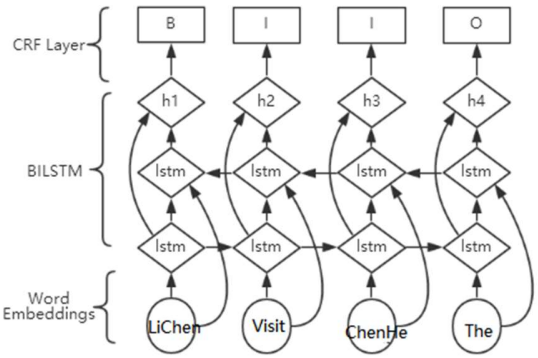

FIGURE 1. Bidirectional LSTM-CRF structure

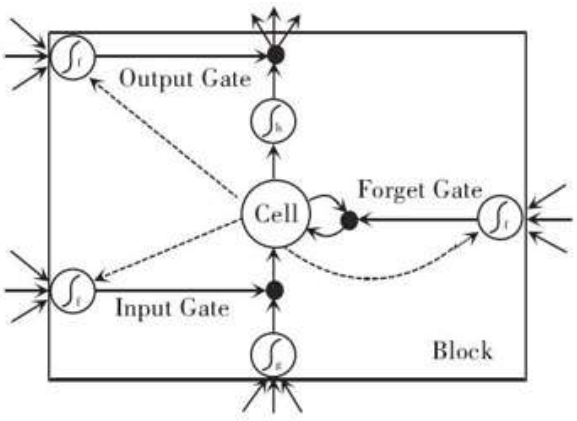

FIGURE 2. LSTM cell

Input gate. The input parameters are the input information of the current position and the information transmitted by the previous hidden neuron. The role of the input gate is to determine the input information, leaving the required ones and cutting down the useless ones.

Forget Gate, which is used to determine how much information the hidden neuron of the previous layer needs to retain and how much to delete.

The Output Gate that determines which of the final output messages are generated and which are not.

From the above description, for the traditional RNN, because the model shares a set of parameters, it is not possible to determine where information is more important and where information is not important, so it is more difficult to learn, and LSTM introduces. After this kind of mechanism, it will be easier to retain important information that the institute needs. And to a certain extent, it can eliminate the problem of the disappearance of gradients.

Based on the LSTM model, this paper uses a bidirectional LSTM and CRF layer structure, as shown in Figure 1. Its interpretation can be interpreted as follows:

Bidirectional LSTM. In sequence labeling tasks, it is often necessary to consider both historical and future contextual information. However, the LSTM's hidden layer unit only records historical information and has no knowledge of future information. The bidirectional LSTM model can be used to solve this problem.

CRF layer. According to existing researches, the neural network structure can be used as a feature extractor, CRF as an outer layer decoding structure to use CRF to model the sequence the objective function of the model is the same as the CRF, except that the feature $h_m\ (s_{t-1}, s_t, , l_o^{t+d})$ of the model is learned through the RNN network structure. For this purpose, the characteristics can be divided into the transition characteristics $h_p\ (s_{t-1}, s_t)$ and the label characteristics $h_q(s_t, , l_o^{t+d})$, Learn these two types of features through the RNN network. So, the objective function of CRF can be rewritten into:

$$\mathrm{H}(s_{t-1}, s_t, l_{t-d}^{t+d}) = \sum_{m=1}^{M} \lambda_m h_m(s_{t-1}, s_t, l_o^{t+d}) = \sum_{p=1}^{P} \lambda_p h_p(s_{t-1}, s_t) + \sum_{q=1}^{Q} \lambda_q h_q(s_t, l_o^{t+d})$$

In the traditional CRF, the characteristic $h_m\ (s_{t-1}, s_t, , l_o^{t+d})$ is usually a 0-1 dispersion value, so the goal to be learned is the weight λ_m. In the neural network structure, $h_q(s_t, , l_o^{t+d})$can be continuous values and updated by reciprocal propagation.

Revise the Results by POS

In traditional grammar, a POS is a category of words that have similar grammatical properties. Words that are assigned to the same POS display similar behavior in terms of syntax. And they play similar roles within the grammatical structure of sentences. Each sentence has its own grammatical structure. We can use POS to get sentences with correct grammatical structure.

The results learned by bidirectional LSTM-CRF Model may not be a complete sentence. For example, as is shown in Figure.3, before we use POS, we get the result “Susan suddenly pregnant and” and we get the corret result“Susan suddenly pregnant” after we use POS to modified the result.

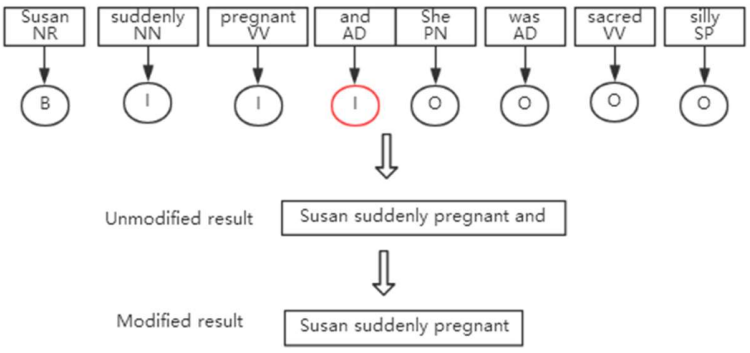

FIGURE 3. Example for POS

EXPERIMENTS AND ANALYSIS

Data Preprocessing

To verify our method, we use 10000 article titles of WeChat public number. Two-thirds of the dataset is used as a training dataset, and one-third of the data is used as a test dataset. This paper uses the jieba word segmentation tool to segment the text and filter out the stop words. We the open source tool word2vec [4] to train word vector proposed by Google in 2013, the selected model is the skip-gram, the dimension is 50, the window size is 5, and the word vector is dynamically updated during the training process. We use the Stand ford CoreNLP toolkit to obtain the POS.

Experimental Parameter Settings

The rest of the parameters that need to be trained are initialized with an even distribution (-0.1, 0.1). The maximum number of words is 20, and the dimension of the input word vector is 50 dimensions. In order to accelerate the training of the model, a batch training method is used, the batch size is set to 256, and the filling method is used to ensure that the length of each batch of training sentences is the same, and the learning rate is Set to 0.01, the number of nodes in the hidden layer is set to 128, and λ is set to 0.01. Adagrad [5] optimizer is used to train the model.

Chen He teased Jia Ling for her flabby tum, Jia Ling beautiful conterattack you are pussy!	Chen He teased Jia Ling for her flabby tum	1
16 years ago, Fan Bingbing, 20 years old, was amazing. She was very beautiful!	Fan Bingbing, 20 years old, was amazing	1
Sha Yi asked Hu Ke to open his video to his son. His son ask, ″Who?″. The family was very happy.	Sha Yi asked Hu Ke to open his video his son	0
Song Xiaobao wears women's clothes, amazing	Song Xiaobao wears women's clothes	1
Cai Yiling made a video to support Luo Zhixiang, if there is pure friendship in entertainment circles , these stars will tell you the truth.	Cai Yiling made a video to support Luo Zhixiang	1

FIGURE 4. Examples of results

Experimental Results and Analysis

We randomly selected 445 data tags in the test data set, 1 indicates that the generated topic summary is qualified, and 0 indicates that the generated topic summary is not qualified. Among them, 377 is the qualified results, and the qualification rate is 84.7%. As is shown in Figure.4, The summary generated by the method of this article is more in line with the grammar rules and the sentence pattern is more complete.

CONCLUSION

In this work, we propose a bidirectional LSTM-CRF model for article title summarization and use POS to modified the results generated by the model bidirectional LSTM-CRF. The results show that our method is effective. As future work, we plan to add POS features to input layer and apply our method to long text summarization task.

ACKNOWLEDGMENTS

First and foremost, we would like to show my deepest gratitude to my mentor Mr. Wei, who has provided me with valuable guidance in every stage of the experiment, and colleague Mr. Wang for insightful discussions.

REFERENCES

1. Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv preprint arXiv:1508.01991, 2015.
2. Lample G, Ballesteros M, Subramanian S, et al. Neural architectures for named entity recognition[J]. arXiv preprint arXiv:1603.01360, 2016.
3. Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.
4. Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]//Advances in neural information processing systems. 2013: 3111-3119.
5. Duchi J, Hazan E, Singer Y. Adaptive sub gradient methods for online learning and stochastic optimization[J]. Journal of Machine Learning Research, 2011, 12(Jul): 2121-2159.