Research on Construction Method of Chinese NT Clause Based on Attention-LSTM

Teng Mao, Yuyao Zhang, Yuru Jiang(✉), and Yangsen Zhang

Institute of Intelligent Information Processing,
Beijing Information Science and Technology University, Beijing, China
maoteng_mt@163.com, yurujiang@123.com

Abstract. The correct definition and recognition of sentences is the basis of NLP. For the characteristics of Chinese text structure, the theory of NT clause was proposed from the perspective of micro topics. Based on this theory, this paper proposes a novel method for construction NT clause. Firstly, this paper proposes a neural network model based on Attention and LSTM (Attention-LSTM), which can identify the location of the missing Naming, and uses manually annotated corpus to train the Attention-LSTM. Secondly, in the process of constructing NT clause, the trained Attention-LSTM is used to identify the location of the missing Naming. Then the NT clause can be constructed. The accuracy of the experimental result is 81.74% (+4.5%). This paper can provide support for the task of text understanding, such as Machine Translation, Information Extraction, Man-machine Dialogue.

Keywords: NT clause · Attention-LSTM · Text understanding

1 Introduction

In Natural Language Processing (NLP), many tasks centers on text analysis, such as Machine Translation, Information Extraction and so on. The basic unit of a text is *sentence*, so the correct definition and cognition of *sentence* is very important for NLP.

How to define a sentence? Bloomfield proposed "Any sentence is an independent form of language, and should not be included in any larger form of language by any grammatical structure. The sentence can be divided by this fact." [1]. He emphasized the independence of sentences and the maximum of inclusion. For Indo-European languages, the basic pattern of sentences can be summed up as "subject-predicate" by virtue of whose linguistic characteristics [2]. Relatively speaking, there is no formal rule defining a Chinese *sentence* yet. Despite the odds, scholars have been trying to study Chinese sentences.

Zhu [3] defined sentence in such a way: "There is a pause before and after the sentence, and intonation represented the relative complete meaning of the language form". In the process of sentence cognition, pause and intonation are partially verifiable, but "relatively complete meaning" lacks operational standards. When Xing [4] annotated complex sentences, the principle was that full stop, sign and question mark separated sentences. However, the use of these punctuations in real texts is often

M. Zhang et al. (Eds.): NLPCC 2018, LNAI 11109, pp. 340–350, 2018.
https://doi.org/10.1007/978-3-319-99501-4_30

arbitrary and lacks the constraints of linguistic norms. And, such sentence is incomplete in structure and meaning.

Cao [5] proposed the concept of Topic Chain, and defined that a topic can be shared by more than one sub-clauses. He regarded topics and subjects as two coexisting elements in a text. So the concept of his topic is narrower, and the topic chain structure can't cover the full text of Chinese texts.

Song [6], from the microscopic topic level, put forward the concept of Naming and Telling, defined the concept of Naming Structure, and formed NT clause theory. The theory of NT clause reveals the organization form of Chinese text in micro topic level, and proved its high coverage and operability in a number of corpus.

Based on the theory of NT clause, this paper uses an attention-LSTM model to identify Naming of punctuation sentences to construct NT clause in the text, which provides support for further text understanding.

2 NT Clause Theory

2.1 Basic Concepts

Punctuation Sentence (PS): It is a string of words, which is separated by commas, semicolons, periods, exclamation marks and question marks from text.

Example 1（钱钟书《围城》Zhongshu Gao *Fortress Besieged*)

①高松年发奋办公，②亲兼教务长，③精明得真是睡觉还睁着眼睛，④戴着眼镜，⑥做梦都不含糊的。

(Songnian Gao works hard, who is the chief of the dean of teaching, and is so shrewd that he still opens his eyes, wears glasses and can't be vague when sleeping.)

According to the definition of PS, there are 6 PSs in Example 1. Although PSs are not necessarily independent in structure and meaning, it has pause and intonation, which is the basic element of the sentence. The structure of the sequence of PS is constrained by grammar, so PS has a certain grammatical meaning. PS can be applied to the full text without ambiguity, so it is suitable for machine processing. All in all, PC is the basic unit of NT clause.

Naming-Telling: In the context of a text, if one component in PS is mentioned by other PS, the former is called the latter's Naming, and the latter is the former's Telling. Each Naming governs one or more PS as its Telling. The structure of this Naming-Telling relation is called the Naming Structure (NS). Using newline-indent schema to represent the header, each PS occupies one line, and indents to the right, and shrinks to the right side of the upstream Naming. Example 1 can be represented as Fig. 1.

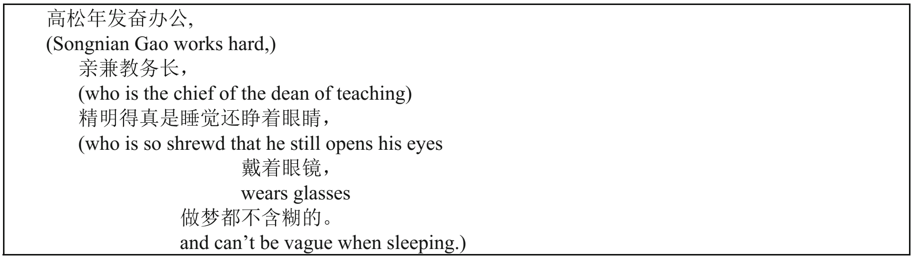
高松年发奋办公，
(Songnian Gao works hard,)
亲兼教务长，
(who is the chief of the dean of teaching)
精明得真是睡觉还睁着眼睛，
(who is so shrewd that he still opens his eyes
戴着眼镜，
wears glasses
做梦都不含糊的。
and can’t be vague when sleeping.)

Fig. 1. The representation of Example 1

NT Clause: In many PSs, the Naming is missing and the propositional elements are incomplete. After a PS has been supplemented to the component that appears in context and should be used as a Naming of this PS, it is called as Naming Sufficient Clause (NSC). NSC consists of Naming and Telling, therefore it is also called NT clause. If a PS’s head is not missing, it is called a Naming Structure Independent Sentence (NSIS), also referred to as an Independent Sentence (IS). In Example 1, ① is a IS. The NT clause of all PSs in Example 1 is shown as Fig. 2.

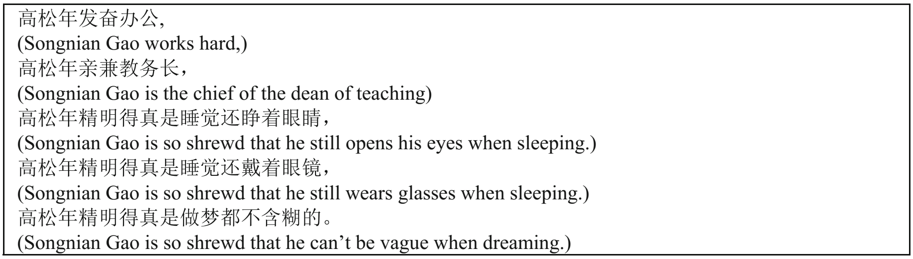
高松年发奋办公，
(Songnian Gao works hard,)
高松年亲兼教务长，
(Songnian Gao is the chief of the dean of teaching)
高松年精明得真是睡觉还睁着眼睛，
(Songnian Gao is so shrewd that he still opens his eyes when sleeping.)
高松年精明得真是睡觉还戴着眼镜，
(Songnian Gao is so shrewd that he still wears glasses when sleeping.)
高松年精明得真是做梦都不含糊的。
(Songnian Gao is so shrewd that he can’t be vague when dreaming.)

Fig. 2. NT clauses of all PSs in Example 1

2.2 Construction NT Clause

In the case of human understanding, they can recognize the missing words of each PS in turn, that is to say, they can obtain NT clauses from PSs. This process relies on the semantic relationship between words and PSs, and it is also restricted by formal patterns. How does the machine carry out the same process?

In previous studies, some traditional methods have been used to construct NT clause [7–10], and the accuracy of single PS is 77.24% [10], sequence PS is 67.37% [10]. In recent years, the neural network has developed and perfected gradually. It has strong representation ability, can obtain grammar and semantic information in the context and has good performance on many Natural Language Processing tasks. Long short-term memory (LSTM) networks is a special form of recurrent neural networks (RNNs) [11]. It can process sequence data and obtain context information of data. Neural Attention Model is first used for machine translation. Bahdanau et al. utilized an attention-based contextual encoder to constructs a representation based on the generation context [12].

The core goal of Attention is to select more critical information among many information sources for the goal of current task. Therefore, this paper constructs a neural network model based on Attention and LSTM to construct NT clause.

3 Method of Construction NT Clause Based on Attention-LSTM

In this paper, the Chinese word is represented as w, and the PS in the text is represented as $c = \{w_1, w_2, \ldots, w_i, \ldots, w_m\}$, where w_i is the i^{th} word in c and m is the number of words in c. The NT clause of c is represented as $t = \left\{w_1, w_2, \ldots, w_j, \ldots, w_l\right\}$, where w_j is the j^{th} word in t and l is the number of words in t. The PS sequence is represented as $C = \{c_1, c_2, \ldots, c_i, \ldots, c_n\}$, where c_i is the i^{th} PS in C and m is the number of words in C, it is defined that $c_{i,i}$ is the i^{th} word w_i of PS c_i in C, the default is c_1 is an Independent Sentence. The NT clause sequence corresponding to the PS sequence is represented as $T = \{t_1, t_2, \ldots, t_i, \ldots, t_n\}$, where the default is $c_1 = t_1$, and $t_{i,j}$ is the j^{th} word w_j of the NT clause t_i in T.

In order to construct NT clause, this paper proposes a model based on Attention-LSTM, which can identify the location of the missing Naming. This task can be divided into two tasks: the first task is to construct the NT clause of a single PS; the second task is the dynamic construction of the NT clause sequence for sequence PS. In the first task the trained Attention-LSTM is used to identify the location of the missing Naming. Then the NT clause of a single punctuation sentence can be constructed. This paper focuses on the first task.

3.1 Attention-LSTM Model

In this paper, the input of Attention-LSTM is the NT clause t_i of PS c_i and the next PS c_{i+1} in the text, in which i represents the position of the PS in text, the target output $index\_pred_i$ is the location of the missing Naming of the c_{i+1} in t_i, which can be represented as $AttentionLSTM(t_i, c_{i+1}) = index\_pred_i$. The framework of Attention - LSTM model constructed is shown in Fig. 3, which has four parts: Word Embedding, Contextual Embedding, Attention -LSTM Layer, Output Layer.

Word Embedding maps each word in t_i and c_{i+1} to a hight-dimensional vector space. This paper uses pre-trained word vectors to obtain the fixed word embedding for each word. The output of Word Embedding are two matrices: $T \in R^{m \times d}$ for t_i and $C \in R^{n \times d}$ for c_{i+1}, where d is the dimension of word vector.

Contextual Embedding uses a BiLSTM on top of Word Embedding to obtain contextual information from surrounding words, so that if refines the embedding of the words. Hence $H\_T \in R^{m \times 2d}$ can be obtained from $T \in R^{m \times d}$, and $H\_C \in R^{n \times 2d}$ from $C \in R^{n \times d}$, where column is $2d$-dimensional because of the concatenation of the outputs of the forward and backward BiLSTM, each with d-dimensional output. Note that in order to ensure consistent representation performance, the same BiLSTM is respectively used in T and C.

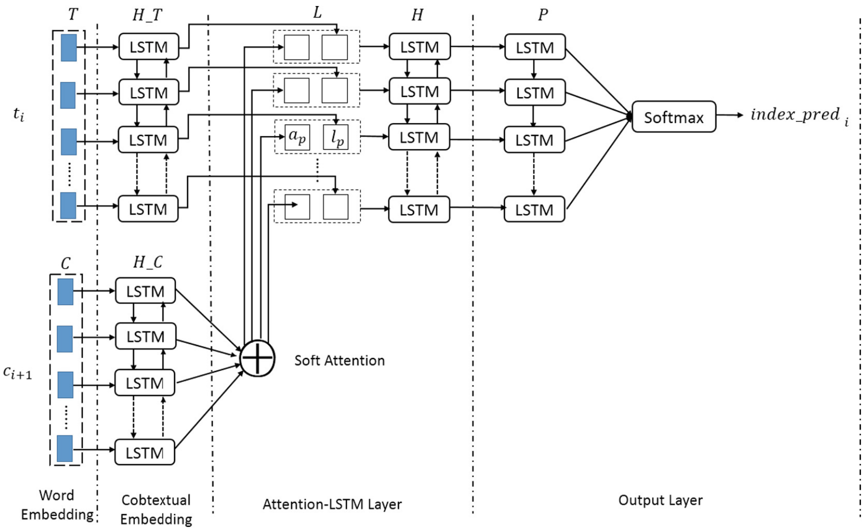

Fig. 3. The architecture of attention-LSTM model

Attention-LSTM Layer tries to match t_i against c_{i+1}. At position p of t_i, it first uses the standard word-by-word attention mechanism [13] to obtain attention weight vector: $a_p \in R^n$. a_{pq} indicates the degree of matching between the p^{th} token in t_i with the q^{th} token in c_{i+1}. Then this paper uses the attention weight vector: a_p to obtain a weighted version of H_C and combine it with the current token of t_i to form a vector $l_p \in R^{2d}$. After each token in t_i is processed, $L \in R^{m \times 2d}$ is obtained, and fed into a BiLSTM.

After this BiLSTM, $H \in R^{m \times 4d}$ is obtained, where column is $4d$-dimensional because of the concatenation of the outputs of the forward and backward BiLSTM, each with $2d$-dimensional output.

The input to Output Layer is H, which encodes the match information between t_i and c_{i+1}. H is fed into LSTM, and the expected output is the probability matrix of *index* at each location: $P \in R^m$, where p_k is the probability of *index* as k. Finally, P is fed into Softmax, the result is $index\_pred_i$.

3.2 Construction of NT Clause for a Single PS

When constructing NT clause for a single PS, the input is the NT clause t_i of PS c_i and the next PS c_{i+1} in the text, in which i represents the position of the PS in text, the target output is the NT clause t_{i+1} of PS c_{i+1}. This paper first uses the Attention - LSTM model to predict position: $index\_pred_i$, and then generates NT clause: $t\_pred_{i+1} = \left\{t_{i,1}, .., t_{i,index}, c_{i+1,1}, \ldots, c_{i+1,m}\right\}$. For example, Example 1 can be divided into 6 processes of constructing NT clauses for a single PS, such as t_1: {高松年发奋 办公(Songnian Gao works hard)} and c_2: {亲兼 教务长(who is the chief of the dean of teaching)} as input, target output is $index\_pred_2 = 1$, t_pred_2 = {高松年亲兼 教务长(Songnian Gao is the chief of the dean of teaching)}.

4 Experiments

4.1 Dataset

The data used in this paper is a big sequence PSs with 11,790 PSs, which is derived from the Encyclopedia of fish. After manual annotation, it has already formed a sequence of NT clauses. When constructing NT clause, data can be represented in triples: $(t_i, c_{i+1}, index_i)$, where t_i is NT clause of c_i, c_{i+1} is next PS of c_i, *index* is align position. This paper transforms the original data into triples of this format, and a total of 11789 data are obtained, which is called All-DATA. Part of All-DATA is shown as follows.

	t_i	c_{i+1}	$index_i$
1	鮟鱇目 是 硬骨鱼纲 的 1 目 ，	有3 亚目 16 科 64 属 265 种 。	1
2	鮟鱇目 有 3 亚目 16 科 64 属 265 种 。	因 胸鳍 形成 假臂，	1
3	鮟鱇目 因 胸鳍 形成 假臂，	在 旧 的 系统 中 称为 柄鳍类 。	1
4	鮟鱇目 在 旧 的 系统 中 称为 柄鳍类 。	体 平扁 或 侧扁 ，	1
5	鮟鱇目 体 平扁 或 侧扁 ，	粗短 或 延长 。	2
6	鮟鱇目 体 粗短 或 延长 。	头 大 ，	1
7	鮟鱇目 头 大 ，	平扁 或 侧扁 。	2

In All-Data, this paper carries out a statistical analysis of the number of data corresponding to different *index*. The statistical results are shown in Fig. 4. It can be seen that the *index* range is 0–24 in All-Data. Note that $index = 0$ means that c_{i+1} is an Independent Sentence. The distribution of *index* is very uneven, of which $index = 1$ has the largest number of data, and *index* is mainly distributed between 0–10, accounting for 99.14% of the total data. Except for 0, with the increase of *index*, the number of data is getting smaller and smaller.

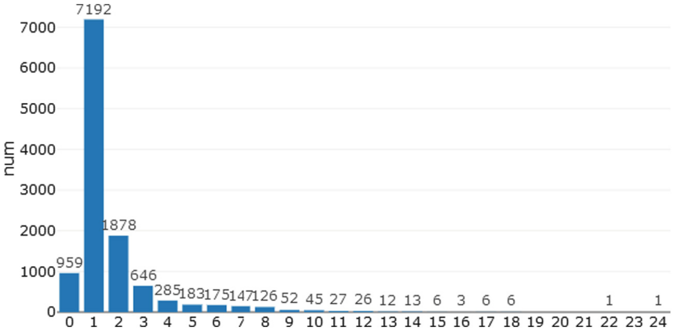

Fig. 4. Distribution of *index* in all-dataset

The training of Attention-LSTM requires a large number of data. Therefore, this paper divides the All-Data according to the proportion of 9:1. 9/10 of All-Data as ALSTM-Dataset, which is used to train Attention-LSTM model. 1/10 of All-Data as

Test-Dataset to do experiment on the construction of NT clause. In addition, the distribution ratios of different *index* in different dataset are calculated respectively, and the results are shown in Fig. 5.

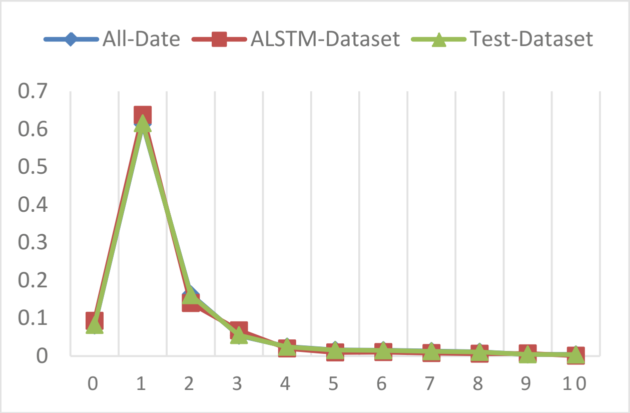

Fig. 5. Distribution of *index* in different dataset

It is can be seen that the lines of All-Data and ALSTM -Dataset almost overlap, and have little difference compared with Test-Dataset's. Therefore, the distribution ratio of *index* is almost the same in different datasets, and the three datasets can be regarded as undifferentiated datasets.

4.2 Attention-LSTM Model Details

All sequence in All-Dataset are tokenized by Jieba[1], a segmentation tool Pre-trained word vectors used in this paper is trained by word2vec module in gensim[2] from BaiduBaike croups[3], the dimension d is fix to 200. This paper uses the AdaDelta [14] optimizer, with a minibatch size of 64 and an initial learning rate of 0.5, for 30 epochs. A dropout [15] rate is 0.3 in LSTM and BiLSTM.

4.3 Training on Attention-LSTM

For training the Attention-LSTM model, this paper divides ALSTM-Dataset into Train-Data and Valid-Data according to the proportion of 9:1, in which Train-Data is used to train model, and Valid-Data is used to verify the model performance. In the training process, the result of each epoch is shown in Fig. 6, where line of Train-Acc almost overlap with Train-F1. In 0–5 epoch, Acc_{index} and F_1-*Score* of Train-Data and Valid-

[1] https://pypi.python.org/pypi/jieba/.

[2] https://radimrehurek.com/gensim/moels/word2vec.html.

[3] https://baike.baidu.com/.

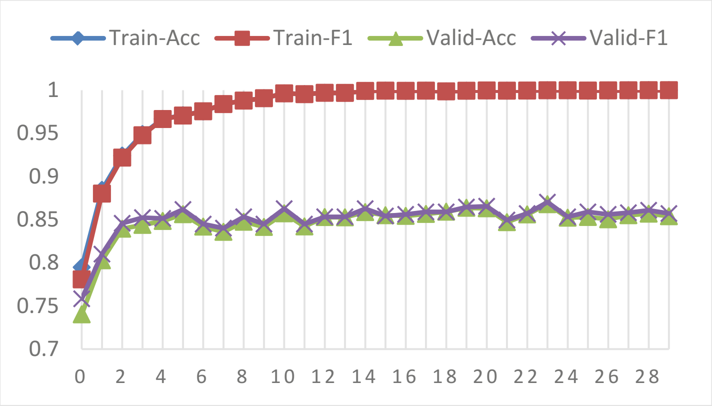

Fig. 6. Training process of model

Data rapidly increased. In 5–29 epoch, the results on the Train-Data is obviously better than Valid-Data, the phenomenon of over fitting appeared.

In order to verify the stability and reliability of the model, 10-fold cross-validation is carried out. The SNT-Dataset data is divided into 10 parts: 9 of them as training data, the remaining as the test data, then the mean value of the 10 results is regarded as the accuracy of the model. The results of 10-fold cross-validation are shown in Fig. 7.

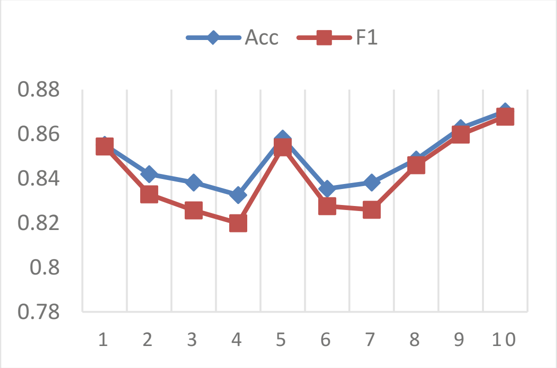

Fig. 7. Results of 10-fold cross-validation in first task

From Fig. 7, we can see that the maximum difference between results is 0.05%, and the fluctuation is small. Therefore the model has reliable stability.

4.4 Experiment on the Construction of NT Clause for a PS

In the process of constructing single NT clause for a PS, This paper separately uses 10 trained models, which were trained in the 10-fold cross-validation process to construct NT clause for 1,158 punctuation sentences in Test-Dataset. The results are shown in Fig. 8:

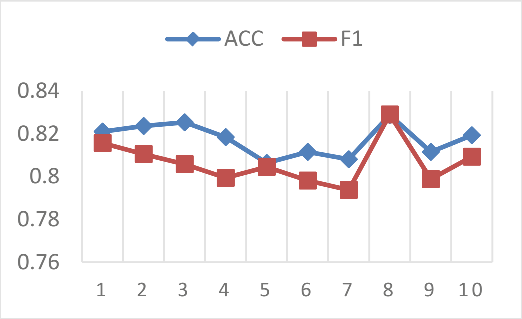

Fig. 8. Results of 10-fold cross-validation in first task

From Fig. 8, we can see that the maximum difference between results is 0.04%, and the fluctuation is small. Therefore the model has reliable stability. Finally, Acc_{index} is 81.74% and F_1-*Score* is 80.65%.Acc_{index} is improved by 4.5%.

This paper also analyzes the model performance in different *index* data, as shown in Fig. 9 (some *index* are not listed in the diagram because they did not appear in the prediction results).

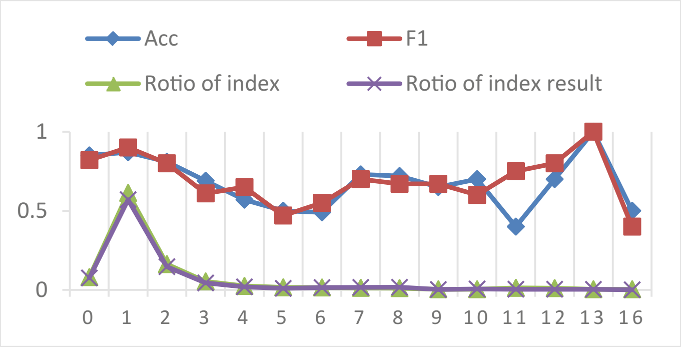

Fig. 9. Results in different *index* in first task

As can be seen from Fig. 9, *index* distribution of the result is basically the same as that in All-Data, which indicates that although *index* distribution is unbalanced, the effect on the model is less. The results of different *index* are closely related to the number of data. In general, the more data there is, the higher accuracy it shows, such as 0–9. However, when the data is small, the results are fluctuating and not reliable enough, such as 10–24.

5 Conclusion

This paper briefly describes the theory of NT clause and proposes a novel method based on Attention-LSTM model for construction NT clause. In the experiment, dataset is taken from Fish BaiduBaike corpus, and it is extracted after the data is manually labeled and processed, then it is used to train Attention-LSTM model. The results show that the method has certain advantages.Acc_{index} is 81.74%(+4.5). However, there are some deficiencies in this research, because the performance of neural network model depends on the quantity and quality of training data, and the number of different *index* data is different in this paper. Therefore, the performance of the model varies greatly on different indexes. Future work in this same direction of study is to add more features into the model, such as POS, grammar and so on, so as to reduce the difference on different *index*. All in all, NT clause construction based on NT clause theory can benefit other text understanding tasks, such as Machine Translation, Information Extraction, Man-machine Dialogue.

Acknowledgements. This work was supported by grants from National Nature Science Foundation of China (No. 61602044), National Nature Science Foundation of China (No. 61370139), Scientific Research Project of Beijing Educational Committee (No. KM201711232022).

References

1. Bloomfield, L.: Language. George Allen and Unwin, American (1933)
2. Jianming, L.: Characteristics of Chinese sentences. Chin. Lang. Learn. **1**, 1–6 (1993)
3. Dexi, Zh.: Grammar Lecture. The Commercial Press, China (2003)
4. Fuyi, X.: Research on Chinese Complex Sentences. The Commercial Press, China (2001)
5. Fengpu, C., Wang, J.: Sentence and Clause Structure in Chinese: A Functional Perspective. Beijing Language and Culture University Press, China (2005)
6. Rou, S.: Chinese Clause Complex and the Naming Structure. Empirical and Corpus Linguistic Frontiers. China Social Sciences Press, China. (To be published, 2018)
7. Yuru, J., Rou, S.: Topic clause identification based on generalized topic theory. J. Chin. Inf. Process. **26**(5), 114–119 (2012)
8. Yuru, J., Rou, S.: Topic Structure Identification of PClause Sequence Based on Generalized Topic Theory. Natural Language Processing and Chinese Computing, pp. 85–96 (2012)
9. Yuru, J., Rou, S.: Optimization of candidate topic clause evaluation function in topic clause identification. J. Beijing Univ. Technol. **40**(1), 43–48 (2014)
10. Yuru, J., Rou, S.: Topic clause identification method based on specific features. J. Comput. Appl. **34**(05), 1345–1349 (2014)
11. Sepp, H., Jürgen, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)

12. Dzmitry, B., Cho, K., Yoshua, B.: Neural machine translation by jointly learning to align and translate. CoRR, abs/1409.0473 (2014)
13. Tim, R., Grefenstette, E., Hermann, K.M., et al.: Reasoning about entailment with neural attention (2015)
14. Zeiler, M.D.: AdaDelta: an adaptive learning rate method. arXiv preprint arXiv:1212.5701 (2012)
15. Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. JMLR (2014)