Chinese and English Elementary Discourse Units Recognition based on Bi-LSTM-CRF Model

Yancui Li **Chunxiao Lai** **Jike Feng** **Hongyu Feng**
School of Information Engineering, Henan Institute of Science and Technology,
Xinxiang , Henan, China
Key Laboratory of Advanced Theory and Application in Statistics and Data Science
(East China Normal University), Ministry of Education,
Shanghai, China
Liyancui@hist.edu.cn

Abstract

Elementary Discourse Unit (EDU) recognition is the basic task of discourse analysis, and the Chinese and English discourse alignment corpus is helpful to the studies of EDU recognition. This paper first builds Chinese-English parallel discourse corpus, in which EDUs are annotated and aligned. Then, we present the framework of Bi-LSTM-CRF EDUs recognition model using word embedding, POS and syntactic features, which can combine the advantage of CRF and Bi-LSTM. The results show that F1 is about 2% higher than the traditional method. Compared with CRF and Bi-LSTM, the Bi-LSTM-CRF model can combine the advantages of them and obtains satisfactory results for Chinese and English EDUs recognition. The experiment of feature contribution shows that using all features together can get best result, the syntactic feature outperforms than other features.

1 Introduction

Discourse analysis is helpful for the performance of machine translation, question answering, summarization and other application. EDU recognition is a basic work in discourse analysis task. Only by recognition EDU, can we make further discourse analysis or other works. At present, the existing Chinese-English parallel corpus only align paragraphs, sentences and other linguistic units, but do not annotate bilingual EDUs alignment, which due to EDU recognition is mainly carried out on monolingual. However, EDUs recognition on Chinese and English is vital to bilingual analysis, machine translation et al. For Example1 is a bilingual sentence of Chinese and English, Chinese EDUs are numbered sequentially by e1, e2 and e3, and English EDUs are marked by e1′, e2′ and e3′. Obviously, e1 and e1′, e2 and e2′, e3 and e3′ are alignment pair.

Example 1 A) [京杭运河古来繁华，] ^{e1}[两岸商贾云集，]^{e2} [贸易发达。]^{e3}

B) [The Beijing - Hangzhou Grand Canal has been prosperous since ancient times,] $^{e1'}$ [with both sides of the bank swarming with merchants] $^{e2'}$ [and well - developed trade.]$^{e3'}$

The main work of this paper is recognition the EDUs of Chinese and English as much as possible. The following is the contribution of this paper:

We annotate Chinese and English discourse alignment corpus, which is first corpus contain EDUs alignment information as far as we know;

We get satisfactory results without any handcraft feature by using Bi-LSTM-CRF Model;

We conduct to find out the contribution of various model and features.

This paper combines existing research and Chinese-English discourse alignment corpus to identify and analyze Chinese-English EDUs. Section 2 builds Chinese-English EDUs alignment corpus; Section 3 describes the Bi-LSTM-CRF model and the framework this paper used; Section 4 reports and analyzes the experimental results; Section 5 overviews the related work; Finally, Section 6 summarizes this paper and points out the future research direction.

Proceedings of the 19th China National Conference on Computational Linguistics, pages 1069-1079, Hainan, China, October 31 - Novermber 1, 2020.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

2 Chinese-English Alignment Corpus

2.1 Chinese-English EDUs alignment methods

In order to represent the discourse, the first task is to define the EDUs. Inspired by the work of Li et al(2014) and Feng (2013), we give the definition of Chinese EDUs. Firstly, a clause should contain more than one predicate, expressing not less than one proposition. Secondly, one EDU should have propositional function to another EDU. Finally, a clause should be segmented by some punctuation. As for English EDU, it is the corresponding content of Chinese EDU.

When annotate, we dividing Chinese sentence into parts, and adapting the alignment strategy of the source language is preferred. That is to say, it is segmented according to the established Chinese EDUs, and then align in English. Therefore, EDUs in Chinese and English sentences is correspondence. Such as Example 1, recognition and alignment are achieved under the guidance of this principle. Since Chinese EDUs are preferential when making alignment rules, some sentences with widely ranges may appear in English translation. These EDUs are not adjacent in English sentences, it will affect the alignment of Chinese and English EDUs. EDUs cannot be completely corresponding in this case, and the solution is to align the main parts.

2.2 Chinese-English alignment corpus

According to the alignment annotation principle mentioned in the 2.1. We annotate alignment corpus of Chinese and English. Corpus select from Xinhua daily, and we have marked 100 Chinese-English translation documents. The Chinese-English parallel corpus is marked with Chinese as the main language, supplemented the parallel EDUs by English.

Due to the marked Chinese-English alignment corpus has many contents, and experiments are mainly for EDUs, this paper mainly introduces the annotation principle of EDU in corpus. After practical operation and analysis, the following three points are obtained:

1) The meaning of English and Chinese sentences. According to the logical semantic relations, the corresponding relations of the adjacent EDUs in the alignment corpus can be found respectively, and the relationship is used to divide and align English-Chinese corpus.

2) Structure. Combined with the structure of Chinese language and English language, the order of subject-verb-object in English-Chinese is consistent, and the translation of some noun clauses and adverbial clauses are also consistent, so it is possible to find out the corresponding words in English-Chinese so as to find the corresponding sentence components in English-Chinese for division.

3) Following the punctuation clues. In the translated English corpus, the punctuation in English is mostly consistent with that in Chinese. And according to the distribution of punctuation, the meaning of the text and the translated English EDUs can be more clearly inferred.

There are 100 documents, 513 paragraphs, 899 Chinese sentences,1281 English sentences and 2153 Chinese-English EDU pairs which have been effectively marked. The Chinese EDU average length is 11 words, while the English EDU average length is 20 words. In the paper, the preset program is used to automatically find the parent node information of English EDUs, and the search is carried out in the automatic syntactic analysis tree of Stanford. The method of search is to look up the words from the beginning and the end of the English clauses successively until a common parent node is found in the syntactic tree. By the way of making statistics on the information of parent node which can be found, it is not difficult to find that the main syntactic structure which can make Chinese EDUs corresponding to English clauses are S、VP、NP、PP etc. The syntactic structure and occurrence frequency corresponding to English EDUs are shown in Figure1 From Figure1 we can see most of EDU’s syntactic tag are S and VP, which is consistent with the definition of our EDU.

2.3 Tagging Strategies and Consistency

Two senior students of Chinese department carried out annotation training under the guidance of the project supervisor. 20 parallel paragraphs were randomly selected from the Xinhua daily to mark training corpus. We developed a platform for EDU annotation. The annotation training is mainly composed of three stages: 1) The tutor demonstrates the annotation of 10 documents, and explains the main annotation

Proceedings of the 19th China National Conference on Computational Linguistics, pages 1069-1079, Hainan, China, October 31 - Novermber 1, 2020.

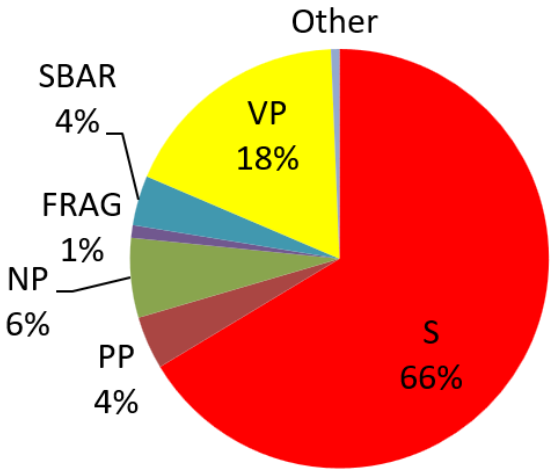

Figure 1: Syntactic structure distribution of the EDUs

strategies, the annotation method and the operation of the annotation platform; 2) Two students mark the remaining 20 documents respectively; 3) Two students respectively proofread the 60 documents marked by themselves with the tutor, and the proofreading was completed in three times, mainly discussing the existing problems and the strategies and methods of correction and annotation. On this basis, the two students annotated the whole corpus together.

In the annotation, we employ left to right segment and alignment method. Consistency is a major criterion of annotation quality. The alignment EDUs annotation evaluation should take into account the recognition consistency and alignment consistency. So, the consistency of Chinese annotation, English annotation, and Chinese-English alignment annotation of the two annotators are considered:

Chinese consistency: the consistency of two annotators on the same Chinese text.

English consistency: the consistency of two annotators on the same English text.

Chinese-English alignment consistency: the consistency of Chinese annotation on the same text by two annotators and the consistency of corresponding English alignment annotation.

We use Method1 and Method2 to compute the consistency.

Method1: computes the consistency of all possible annotations. There are punctuation marks at the recognition positions of Chinese EDUs, and punctuation marks that may be used as recognition marks. The recognition of EDUs in English is not based on punctuation, any space can be calculated as the recognition mark.

Method2: calculating the consistency of intersection (A ∩B) in all (A U B). Sentence Position="X1...X2 | Y1...Y2", calculate the case that A and B mark the same position of recognition. Compared with method 1, this method is more accurate and can unify the evaluation criteria of Chinese and English EDUs recognition.

	Chinese consistency	English consistency	Chinese-English alignment consistency
Method1	0.972	0.992	–
Method2	0.968	0.930	0.909

Table 1: The consistency of EDUs annotation for Chinese and English

As shown in Table 1, recognition alignment shows good consistency, with Chinese alignment up to 0.972/0.968, English alignment up to 0.992/0.930. Even under the strictest circumstances of Method2, Chinese-English alignment up to the consistency rate of 0.909.

It is worth noting under the Method1, English consistency is better than Chinese, with 0.992 >0.972. Under the Method2, Chinese better than English, this is because the consistency in the calculation, Chinese punctuation only for limited computation, but the English is for any Spaces.

Proceedings of the 19th China National Conference on Computational Linguistics, pages 1069-1079, Hainan, China, October 31 - Novermber 1, 2020.

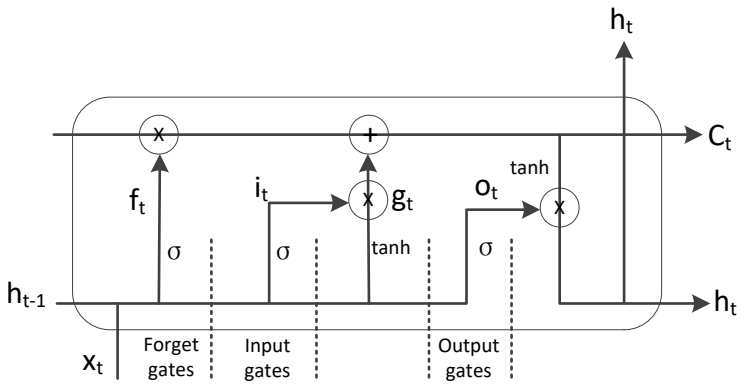

Figure 2: LSTM memory cell

However, the reality is that Chinese alignment is better than English with the same alignment evaluation criteria. This can be shown under Method2 (0.968 >0.930), because Chinese recognition is marked by punctuation, which is relatively easy. However, English recognition is not marked by punctuation, and it is easy to recognition incorrectly. Therefore, Method2 can more accurately reflect the difference in bilingual alignment effect compared with Method1.

3 The Model of EDU Recognition based on Bi-LSTM-CRF

In this section, we introduced the Bi-LSTM-CRF model we used, which is the combination of CRF and Bi-LSTM and have been used in several NLP task.

3.1 CRF

CRF is extension of both Hidden Markov Models and Maximum Entropy Model (Lafferty et al., 2001). It often solves some NLP problems, such as word recognition and image recognition. EDUs recognition is a sequence labeling problem. One solution is that it can assign each word in the sentence with label Y (word is EDU boundary) or N (word is not EDU boundary). CRF is a sequence labelling model with flexible feature space. Therefore, with given feature set and labeled training data, the CRF model solve EDUs recognition task. The model is defined as Eq. (1):

$$p(Y|X) = \frac{1}{Z(X)} exp(\sum_{k} \lambda_k f_k) \tag{1}$$

In Eq. (1), Z(X) is a probability normalization factor conditioned on X. λ_k is the corresponding weight of the feature set. f_k is the input sequence sentences, and Y is the output label of Y or N.

3.2 Bi-LSTM

RNN is a model suitable for sequence data, which uses previous and current state to determine the final output. However, in practical applications, RNN has only short-term memory because the gradient vanishing and exploding problem. Hochreiter and Schmidhuber(1997) propose LSTM network ,a variant of RNN to solve this problem.

Figure2 illustrates a single LSTM memory cell. We can see that it contains input, forget and output gate. The gates determine the current information, in a certain proportion or discarded, transferred to the next moment. Through the gate, LSTM can remove or add information to the cellular state. Therefore, they can solution the data long range dependencies problem.

LSTM memory cell is implemented as the Eq.(2): As shown in Eq. (2), the logistic sigmoid function is denoting as σ . it is the input gate. ct is cell vectors. i_t decides the information will be stored in c_t . ft is forgot gate, and it decides the information can through from the previous cell. o_t is output gate,

Proceedings of the 19th China National Conference on Computational Linguistics, pages 1069-1079, Hainan, China, October 31 - Novermber 1, 2020.

$$
\begin{aligned}
i_t&=(W_{ii}x_t+b_{ii}+W_{hi}h_{(t-1)}+b_{hi})\\
f_t&=(W_{if}x_{t+}b_{if+}W_{hf}h_{(t-1)}+b_{hf})\\
g_t&=\tanh(W_{ig}x_t+b_{ig}+W_{hg}h_{(t-1)+}b_{hg})\\
o_{t=}&(W_{io}x_t+b_{io}+W_{ho}h_{(t-1)}+b_{ho})\\
c_t&=f_tc_{(t-1)}+i_tg_t\\
h_{t=}&o_t\tanh(c_t)
\end{aligned}
\qquad (2)
$$

it decides the information output to the current hidden state h_t. W is the weight matrix, and b is bias vectors of each gate. They are learned during training. $\oplus$ denotes the vector concatenation. For sequence tagging task, Graves and Jürgen utilize a bidirectional LSTM(Bi-LSTM) network. Bi-LSTM is extended on the basis of LSTM, and it contains two difference direction layers. The sequence $\overline{h} = (\overline{h_1 h_2}.....\overline{h_n})$ of the Bi-LSTM layer is obtained past and future input features by the forward and backward LSTM. The LSTM allows more context dependent information than LSTM.

3.3 Bi-LSTM-CRF

We describe our Bi-LSTM-CRF models in details. Figure 3 shows the Bi-LSTM-CRF framework. As we can see from Figure 3, there are input layer, embedding layer, Bi-LSTM layer, CRF layer and output layer. First, words in sentences and their features are vectorized. Secondly, the Bi-LSTM model is fed with feature vectors to learn contextual features from the forward and backward directions. Then, Bi-LSTM output result is input to CRF layer. Finally, the CRF layer predicts the globally optimal clause sequence. In addition, to reduce the influence of overfitting, we add a dropout layer at ends of the Bi-LSTM model.

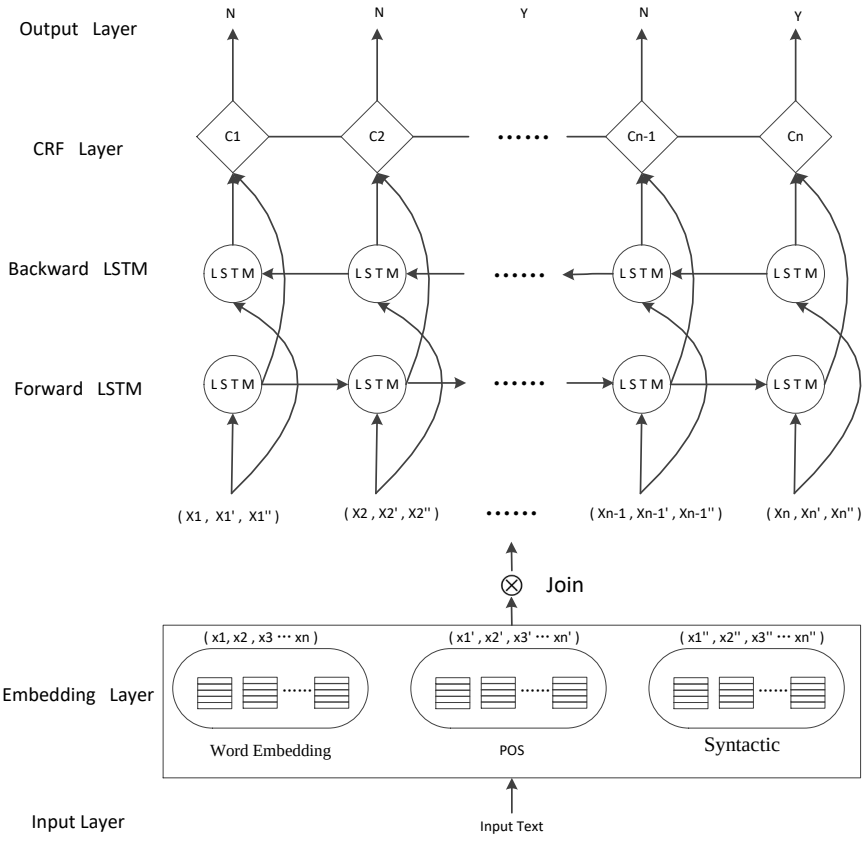

Figure 3: The framework of Bi-LSTM-CRF Model

Bi-LSTM-CRF expands the CRF layer on the basis of Bi-LSTM. The performance of CRF model in sequence annotation tasks has been verified. In this model, the Bi-LSTM through Bi-LSTM layer makes full use of past and future information, and CRF layer make use of tag in-formation. So, this model can predict the current tag by incorporate the advantage of Bi-LSTM and CRF.

4 Experiment Results

4.1 Experiment Setting

The input layers of our models are the input text. We give the vector representations of words, Part of speech (POS) and Syntactic. Word embeddings are pretrained using skip-n-gram, a variation of word2vec (Mikolov et al.,2013) that sensitive to the order of word. These embeddings are adjusted during training. We find improvements using pretrained word embeddings. For English, the embedding dimension we used is 200. For Chinese, we use pre-trained vector files from People's Daily News, the embedding dimension is 300(Li et al,2018). We use dropout training to avoid the model depending on one representation too strongly, and find it is import to result.

POS is the process of marking a word as nouns, verbs, adjectives, adverbs, etc. POS is used in many NLP task and proved very useful. Syntactic is the component that takes input sentence and give the grammatical tree structure of sentence, which is widely used to understanding written language, discourse parsing et al. For example, syntactic can output the phrases tag of words. We use Stanford coreNLP (Mikolov et al.,2013) to get POS and syntactic feature, it can give the POS and syntactic tag of each word. In this paper, we use parent phrase tag as syntactic feature simplify.

The task of EDUs recognition is giving a tag to every word in a sentence. A single EDU could span several words in a sentence. Sentences can represent in the Y(Yes) or N (No) format, where each word is labeled as Y label if the word is the end of EDU, and as N label if it is the beginning or inside of EDU word.

We exploit standard training methods for our model. Using AdaGrad(Duchi,2011) as stochastic gradient decent. Calculate derivatives from standard back propagation (Goller and Kuchler, 2002). In order to prevent over fitting, we regularize our model using dropout method (Srivastava et al.,2014), and fixed rate 0.5 for dropout layer. We obtain improvements after using dropout.

We set the initial AdaGrad learning rate as 0.01. The dimension of pre-trained word embed-ding is set as 200. The dimension of LSTM hidden state as 200.The W and b are randomly initialized with a uniform distribution in the range (-0.01, 0.01). We use publicly available 200-dimensional embeddings trained for English, there are total 40000 words. We use 300-dimensional embeddings for Chinese, there are total 355989 words. The Bi-LSTM units set 256, epoch set 200. In our experiment, for Chinese EDUs recognition, there are total 12 581 words, 32 POS tags and 29 syntactic tags. For English, there are total 4 106 words, 47 POS tags and 29 syntactic tags.

4.2 Experiment Results

In this section, EDU recognition is carried out in our Chinese-English alignment corpus. There are total of 100 documents, 513 paragraphs and 2 153 EDUs were involved. The recognition of English word is 42 122 in total, among which there are 2 153 positive labels. The ratio of positive and negative examples is 19.6:1 for English. Overall, the average length of the English EDUs is about 20 words, while the Chinese EDUs is 11 words. The experiment splits instances into 10 parts, and use 8 parts for training, 1 part for verification and 1 part for testing. The features we used are word embedding, POS tag and syntactic tag. The recognition results of Chinese words boundary are indicated in Table 2.

In Table 2, the best results are highlighted bold for each metric. From Table 2, we can see that by combining Bi-LSTM, pretrained embedding, and CRF on the top of the framework, our Bi-LSTM-CRF model outperforms best of all. We obtain the satisfactory results with the F1 93.4 % and R 94.4 % by using Bi-LSTM-CRF model.

Table 3 shows the English EDUs recognition result. For the purpose of comparison, we list Li' s (Li et al, 2012) Maxent model results together with ours CRF and Bi-LSTM, especially our Bi-LSTM-CRF model results for comparison. The best F1 is 94.4% using Bi-LSTM-CRF model.

Proceedings of the 19th China National Conference on Computational Linguistics, pages 1069-1079, Hainan, China, October 31 - Novermber 1, 2020.

Model	P	R	F1
Li’ s Maxent	87.4	93.6	90.4
CRF	86.7	96	91.1
Bi-LSTM	**95.4**	89.8	92.5
Bi-LSTM-CRF	92.3	**94.4**	**93.4**

Table 2: Chinese EDU words boundary recognition results

Model	P	R	F1
Li’ s Maxent	86.5	78.7	82.4
CRF	87.4	91	89.1
Bi-LSTM	94.0	91.9	92.9
Bi-LSTM-CRF	**95.5**	**93.4**	**94.4**

Table 3: English EDU words boundary recognition results

Figure 4 comprise the result of F1 between Chinese and English for different models. We can see the best model is Bi-LSTM-CRF model, by joint decoding label sequence can benefit the final performance of neural network models, followed by Bi-LSTM and CRF. The reason is that EDUs recognition is sequence tag task, Bi-LSTM and CRF classifier perform better than traditional Maxent classifier.

Figure 4 shows that English EDUs recognition result is higher than Chinese using Bi-LSTM or Bi-LSTM-CRF, the reason is that the pretrained embedding of Chinese words are more than English, with Chinese 35 598 where English 4 000, the two is 10 times difference. But for using Maxent or CRF model, Chinese EDUs identification F1 is higher than English.

4.3 The contribution of features

In order to investigate the contribution of the features, we give experiments specifically targeted at features for EDUs recognition. Table 4 shows the performance of P, R, F1 for Chinese separately using different feature, and Table 5 gives the results of English.

Features	P	R	F1
Word Embedding	65.2	88.6	75.1
POS	70.1	80.2	74.8
Syntactic	81.1	82.1	81.6
Word Embedding +POS	76.7	90.7	83.1
Word Embedding +POS+ Syntactic	92.3	**94.4**	**93.4**

Table 4: The different feature result for Chinese

Table 4 and Table 5 show that syntactic feature outperform than other features, the F1 can reach 81.6% and 81.8% for Chinese and English. The reason is that both in Chinese and English, most EDU word syntactic labels contain IP and VP syntactic, while word with syntactic NP, PP and LCP are not EDU boundary. Syntactic information is highly related with EDUs recognition than other information. The combine of all features performance best both in Chinese and English, that means the more information used, the better the results.

POS is the commonly used in NLP task, from the results, we find it is also useful for EDU recognition. As shown in Table 5, only using word embedding feature, we can get F1 80.4% for English. We also find that word embedding feature is useful than syntactic feature for English, mainly because Chinese word is sparing. And Chinese EDUs boundary usually have punctuation, which have IP tag, so syntactic feature is useful than word embedding feature for Chinese.

According to the results, we know that using word embedding, POS and syntactic feature together, we can get best result, it proves the effectiveness of our features.

Proceedings of the 19th China National Conference on Computational Linguistics, pages 1069-1079, Hainan, China, October 31 - Novermber 1, 2020.

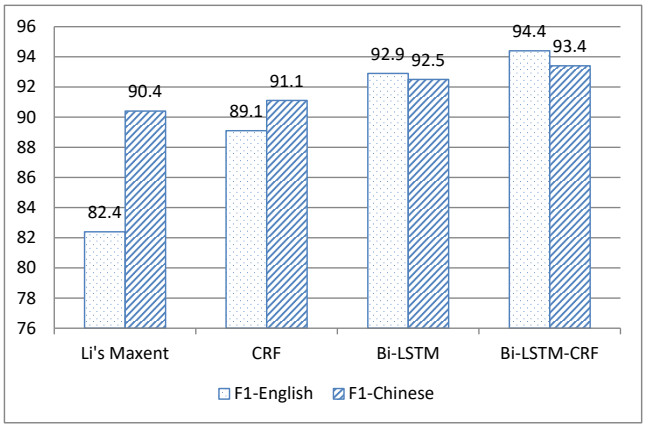

Figure 4: Comparison of F1 between Chinese and English for different models

Features	P	R	F1
Word Embedding	87.4	74.5	80.4
POS	71.2	79.8	75.3
Syntactic	80.2	83.5	81.8
Word Embedding +POS	90.4	87.1	88.7
Word Embedding +POS+ Syntactic	**95.5**	93.4	**94.4**

Table 5: The different feature result for English

4.4 Discussion

There are about 6% EDUs recognition error, and we discuss the reason as follows. There are two cases of errors: one is negative instances are recognized as positive instances. The other is positive instances are recognized as negative instances. From the recognition consistency compute method of section 2.3, we notice the punctuation plays an important role in EDUs recognition, especially in Chinese. For example, if the front words of comma are the subject of the sentence, therefore the position of this comma is not EDU boundary. But when using our model, the syntactic of the words is IP, which may lead to mistake recognition.

In EDUs recognition, it is difficult to distinguish EDUs from complex sentence structure. For example, if you believe that "在…以后，终于(In...After that, finally) "is a connective that expresses the relation of succession. It can be considered as an EDU. However, traditional grammar generally analyzes it as an adverbial, a part of the syntactic structure. This is transition between textual structure and syntactic structure. We currently follow the traditional grammar, leaving the analysis of this situation to the syntactic structure. For the automatic alignment of Chinese and English EDUs, we found that most of EDUs are sequence alignment, only about 4% of EDUs adjusted sequentially when from Chinese to English. So, for EDUs alignment, the main problem is EDUs recognition, which is influence on the result of automatic alignment EDUs. The difficulty of EDUs alignment is that EDUs does not correspond and adjust in order, which needs further research.

This paper only does Chinese and English EDUs recognition respectively, but does not do Chinese-English EDUs alignment. Once EDUs are identified, the next step is to align, and since EDUs are basically one-to-one, EDUs alignment can be turned into a machine translation or classification problem.

Proceedings of the 19th China National Conference on Computational Linguistics, pages 1069-1079, Hainan, China, October 31 - Novermber 1, 2020.

5 Related Work

Due to the emergence of discourse corpus, there have been a lot of researches on the recognition of English discourse. One of the corpora which are widely used is Rhetorical Structure Theory Discourse Treebank (RSTDT) building by Carlson et al. (2003), the other is Penn Discourse Treebank (PDTB) annotated by PDTB Research Group (2007). The RST represents a discourse as a tree, with phrases or clauses as EDU. PDTB adopts the predicate-argument view, with two spans as its arguments.

Due to the EDUs in RST consecutive annotation, the EDUs automatic identification on RSTDT is also called EDUs recognition, and now there is much research on it and the results are ideal, more representative research results include: Soricut and Marcus (2003) adopt statistics method for recognition, the F1 of EDUs recognition on the automatic syntax tree and standard syntax tree are 83.1% and 84.7%. Hernault et al. (2010) give a discourse recognition model based on sequential data annotation. They use lexical and syntactic features get the F1 94%, which is close to 98% of the F1 of manually. According to the above we can know that recogni-tion accuracy of EDUs on RSTDT is relatively high, and there is little room for further improvement. For the un-sequential annotation of arguments on PDTB, not all the discourse is covered. So, some researchers propose to replace the whole argument with the argument center in the recognition of argument (Wellner B. and Pustejovsky J,2007; Elwell R. and Baldridge J.,2008; Wellner B,2009). And other researches put forward to the point of identifying sentences that contain arguments (Prasad et al.,2010), the recognition accuracy of Arg1 and Arg2 are 65% and 85% (Xu F.,2013). Braund et al.(2017) research whether syntax help discourse segmentation, the results show that dependency information is less useful than expected, but they provide a fully scalable, robust model that only relies on part-of-speech information, and show that it performs well across languages in the absence of any gold-standard annotation.

Deep learning method has made breakthroughs in many NLP tasks in recent years. Among them, Cyclic Neural Network (RNN) is a typical sequence marking model, and it is proposed by Goller and Kuchler(1996). However, RNN is limited by gradient disappearance and gradient ex-plosion, Hochreiter and Schmidhuber (1997) come up with the variation of RNN which is named Long Short-Term Memory (LSTM). Because it only gets one-way contextual information, Graves and Schmidhuber (2005) raise the Bi-directional Long Short-Term Memory (Bi-LSTM), and applied it to speech identification. Bi-LSTM can effectively utilize past and future features in a specific time range. On the other hand, Conditional Random Field (CRF) algorithm which is put forward by Lafferty et al.(2001) has been widely applied in NLP recent years. In sequence marking tasks, CRF can take into account the anteroposterior dependence between adjacent la-bels of output. Considering the above reasons, there are some studies attempting to combine Bi-LSTM and CRF to build model for sequence data (Ji Me et al.,2018). Bi-LSTM and CRF hybrid model were first applied to the sequence labeling task of NLP by Huang et al. (2015), Ma and Hovy(2016) focus Bi-LSTM, CRF and CNN models and apply them to sequence marking tasks. Bi-LSTM-CRF model is applied in identifying biomedicine named entity (Greenberg et al.,2018), The effectiveness of the model in sequence marking tasks is gradually verified.

There are few discourse corpora in Chinese to mark EDU information (Zhang et al.2014;Li et al.,2014). At present, the task of EDU recognition is few referred. Zhang et al (2014) only identified the relation, but no relevant result about argument identification. Li et al. (2014) research on Chinese EDUs recognition based on comma, and Chinese EDUs recognition result can reach 90%. Ge Haizhu et al. (2019) proposes a Chinese EDU recognition approach based on theme-rheme theory, which can pay more attention on the internal structure of EDU, and the F1 score is 89.96%. However, limited by bilingual corpus, there is no EDUs recognition of both Chinese and English research.

6 Conclusion

The discourse alignment corpus of Chinese-English is annotated in this paper. The corpus has a complete EDU definition, annotation method, quality assurance and available scale. The corpus we annotated in this paper is the basic task of EDUs recognition. Then we developed an EDUs recognition system using Bi-LSTM-CRF model. Our neural model achieved satisfactory results for Chinese and English EDU recognition. To our knowledge, we are among the first to develop an effective neural network-

Proceedings of the 19th China National Conference on Computational Linguistics, pages 1069-1079, Hainan, China, October 31 - Novermber 1, 2020.

based approach to recognize EDUs for both Chinese and English. We input word embedding, POS and syntactic feature to our model in order to improve the result. By incorporating these features, our model can extract EDUs automatically and high quality. The F1 can reach 93.4% and 94.4% for Chinese and English separately, which is reaching the practical using. This model can also be generalized to solve other problems. In the future, we will improve the effect of recognition Chinese and English EDUs, then try to automatic align them.

Acknowledgements

This paper is supported by the National Natural Science Foundation of China (61502149), by the Open Research Fund of Key Laboratory of Advanced Theory and Application in Statistics and Data Science （East China Normal University）, Ministry of Education(KLATASDS1806), as well as the high-level talent research project of Henan Institute of Science and Technology (2017039).

References

Braud C., Lacroix O., and Anders S. 2017. *Does syntax help discourse segmentation? Not so much.* Conference on Empirical Methods in Natural Language Processing, 2432 – 2442.

Carlson, L.， Marcu, D.， and Okurowski, M. E. 2003. *Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory.* Current and New Directions in Discourse and Dialogue. Springer Netherlands.

Duchi J., Hazan E. , and Singer Y. 2011. *Adaptive Subgradient Methods for Online Learning and Stochastic Optimization.* Journal of Machine Learning Research, 12(7):257-269.

Elwell R. and Baldridge J. 2008. *Discourse connective argument identification with connective specific rankers.* In IEEE International Conference on Semantic Computing, 198-205.

Feng W.H. 2013. *Alignment and Annotation of Chinese-English Discourse Structure Parallel Corpus.* Journal of Chinese Information Processing, 27(6):158-165.

Ge H.Z., Kong F., and Zhou G.D. 2019. *Chinese Elementary Discourse Unit Recognition Based on Theme-Rheme Theory.* Journal of Chinese Information Processing,33(8):20-27.

Goller C., Kuchler A. 1996. *Learning Task-Dependent Distributed Representations by Backpropagation Through Structure.* IEEE International Conference on Neural Networks,347 – 352.

Graves A., Schmidhuber J. 2005. *Framewise phoneme classification with bidirectional LSTM and other neural network architectures.* Neural Networks, 18(5):602-610.

Greenberg N., Bansal T., Verga P., and McCallum A. 2018. *Marginal Likelihood Training of BiLSTM-CRF for Biomedical Named Entity Recognition from Disjoint Label Sets.* In Proceedings of the Conference on Em-pirical Methods in Natural Language Processing, 2824 – 2829.

Hernault H., Bollegala D., and Ishizuka M. 2010. *A Sequential Model for Discourse Recognition.* In Computational Linguistics and Intelligent Text Processing, Springer, Berlin, Heidelberg, 2010, 315-326.

Hochreiter S., Schmidhuber J. 1997. *Long Short-Term Memory.* Neural Computation, 9(8):1735-1780.

Huang Z., Xu W., and Yu K. 2015. *Bidirectional LSTM-CRF Models for Sequence Tagging.* Computation and Language, 2015.

Ji M., Kuzman G. and David W. 2018. *State-of-the-art Chinese Word Recognition with Bi-LSTMs.* In Proceedings of the Conference on Empirical Methods in Natural Language Processing,4902 – 4908.

Lafferty J., Mccallum A., and Pereira F. 2001. *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data.* In Proceedings of the Eighteenth International Conference on Machine Learning , Morgan Kaufmann Publishers Inc, 282-289.

Li S.,Zhao Z. Hu R. et al. 2018. *Analogical Reasoning on Chinese Morphological and Semantic Relations.* In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 138-143.

Li Y.C., Feng W.H., Sun J., et al. 2014. *Building Chinese Discourse Corpus with Connective-driven Dependency Tree Structure.* In proceedings of Empirical Methods in Natural Language Processing, 2105-2114.

Li Y.C., Feng W.H., Zhou G.D., et al. 2013. *Research of Chinese Clause Identification Based on Comma.* Acta Scientiarum Naturalium Universitatis Pekinensis, 49(1):7-14.

Ma X. and Hovy E. 2016 . *End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF.* In Proceedings of the Meeting of the Association for Computational Linguistics, 1064-1074.

Manning C. D., Mihai S., John B. et al. 2014. *The Stanford CoreNLP Natural Language Processing Toolkit.* In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 55-60.

Mikolov T., Sutskever I., Chen K., et al. 2013. *Distributed representations of words and phrases and their compositionality.* Advances in Neural Information Processing Systems, 26:3111-3119.

PDTB Research Group. 2007. *The Penn discourse Treebank 2.0 annotation manual.* IRCS Technical Reports Series.

Prasad R., Joshi A. K., and Webber B. L. 2010. *Exploiting Scope for Shallow Discourse Parsing.* In Proceedings of the Seventh International Conference on Language Resources and their Evaluation,Valletta, Malta, 2076-2083.

Soricut R. and Marcus D. 2003. *Sentence Level Discourse Parsing using Syntactic and Lexical Information.* In Proceedings of the 2003 Conference of the North American,149-156.

Wellner B. and Pustejovsky J. 2007. *Automatically Identifying the Arguments of Discourse Connectives.* In EMNLP-CoNLL, 92-101.

Srivastava N., Hinton G., Krizhevsky A., et al. 2014. *Dropout: A simple way to prevent neural networks from overfitting.* The Journal of Machine Learning Research, 15(1):1929 – 1958.

Wellner B. 2009. *Sequence models and ranking methods for discourse parsing.* Faculty of the Graduate School of Arts and Sciences Brandeis University Computer Science James Pustejovsky, Brandeis University.

Xu F. 2013. *Research of Key Issues in English Discourse Structure Analysis.* Soochow university.

Zhang M.Y., Qin B., and Liu T. 2014. *Chinese Discourse Relation Semantic Taxonomy and Annotation.* Journal of Chinese Information Processing, 28(2):28-36.

Proceedings of the 19th China National Conference on Computational Linguistics, pages 1069-1079, Hainan, China, October 31 - Novermber 1, 2020.