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 1070-1080, Hainan, China, October 31 - Novermber 1, 2020.

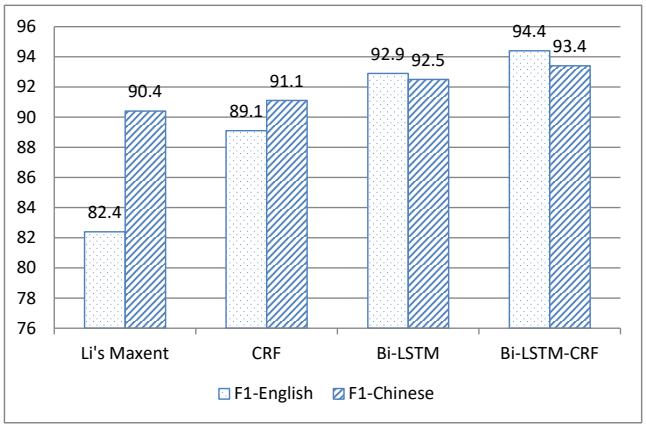

Figure 4: Comparison of F1 between Chinese and English for different models

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 1070-1080, 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 1070-1080, 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 1070-1080, Hainan, China, October 31 - Novermber 1, 2020.