CISA: Chinese Information Structure Analysis for Scientific Writing with Cross-lingual Adversarial Learning

Hen-Hsen Huang[1] and **Hsin-Hsi Chen**[12]
[1] Department of Computer Science and Information Engineering,
National Taiwan University, Taipei, Taiwan
[2] MOST Joint Research Center for AI Technology and All Vista Healthcare, Taipei, Taiwan
hhhuang@nlg.csie.ntu.edu.tw, hhchen@ntu.edu.tw

Abstract

This work demonstrates a writing assistant system that provides high level advice for Chinese scientific writing. Cross-lingual approaches are investigated to analyze the information structure of a given Chinese abstract and retrieve useful knowledge in the related work written in both English and Chinese. To the best of our knowledge, this is the first study on Chinese information structure identification. Without the need of labeled Chinese data, our novel model is capable of dealing with Chinese instances by acquiring language-invariant knowledge from the labeled English data. Adversarial learning is employed to enhance the cross-lingual sentence representation.

1 Introduction

This work presents CISA, a system for Chinese scientific writing advising based on information structure analysis. We extend the scope of our prior system [Huang and Chen, 2017], which addresses English writing advising, to the second largest language in the world. In recent years, a variety of Chinese writing assistant models have been proposed for grammatical error detection and correction [Lee *et al.*, 2016; Shiue *et al.*, 2017]. Beyond the word level and the syntax level, our system focuses on providing the writing advice at the discourse level and the knowledge level.

The core of CISA is a neural network for information structure identification. Once a user submits a scientific article, the function of each sentence will be recognized into five types, including Background, Purpose, Method, Results, and Conclusion. Based on the analysis of writing structure and organization, our system makes writing suggestions and retrieves the related information from a knowledge base (KB). Previous work on information structure identification, or argumentative zoning, has been shown effective in document summarization [Contractor *et al.*, 2012], citation indexing [Teufel, 2006], and literature review [Guo *et al.*, 2014]. However, most of them are limited to English. To the best of our knowledge, there is still no study on Chinese information structure identification, and the corpus has also yet to construct.

Due to the lack of training data, we propose a cross-lingual adversarial network that is trained on the labeled English data and is capable of dealing with Chinese instances. Cross-lingual transfer learning with deep neural network has been shown effective in sentiment analysis [Chen *et al.*, 2016], POS tagging [Kim *et al.*, 2017], and question-answering [Joty *et al.*, 2017]. In our model, cross-lingual information including the Universal POS tags[1] and the bilingual word embedding [Gouws *et al.*, 2015; Vulić and Moens, 2015] are used to represent the sentences, and the sentence representation is further enhanced by using language-adversarial training.

The contributions of this work are three-fold. (1) This work shows the first study on Chinese information structure identification and demonstrates its application for Chinese writing assistant. (2) We propose a novel method that achieves a promising performance without the need of the labeled Chinese data. (3) Our system provides related knowledge for a submitted article with cross-lingual knowledge retrieval. The demonstration system is available online.[2]

2 Method

Figure 1 shows an overview of our model for information structure identification with cross-lingual adversarial learning. The main task is to classify a sentence into one of five types of information structure. The adversarial learning task is to discriminate the language of a sentence. As denoted in Figure 1, the layers in gray are shared between the main network and the discriminator. Our goal is to train a language-neutral sentence representation and force the main network to learn without depending on language-specific information.

The input layer of the sentence representation is the word embeddings initialized with pre-trained bilingual word vectors. We do not update the word embedding layer during training because its cross-lingual capability will be violated when only English data are seen. The sentence representation is encoded by a GRU [Cho *et al.*, 2014] layer.

In addition to the word sequence, the sequence of POS tags, and location features such as the position of the sentence in the abstract are taken into account. Both the Chinese and

[1] http://universaldependencies.org/u/pos/
[2] http://nlg18.csie.ntu.edu.tw/cisa

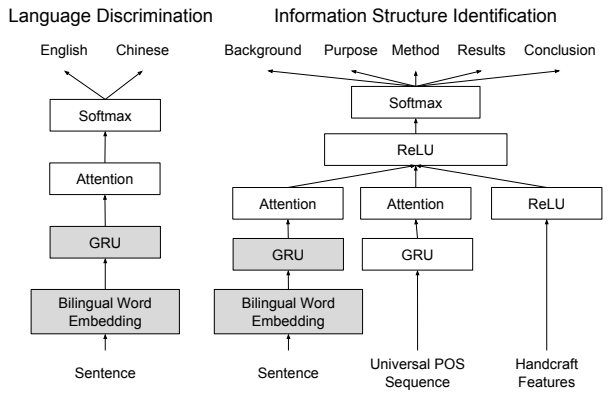

Figure 1: Overview of our model. The layers in gray are shared between the main network and the discriminator.

Type	English	Chinese
Background	471	250
Purpose	590	114
Method	1,117	227
Results	1,031	120
Conclusion	185	25

Table 1: The training (English) and the test (Chinese) data.

the English POS tags are converted to the Universal POS tags for the cross-lingual purpose. Note that language-specific features such as bag of words are excluded although some of them have been shown useful in the monolingual task.

3 Evaluation

NTHU Academic Writing Database,[3] which consists of 597 labeled English abstracts in the EECS domain, is adopted as the training data. For evaluation, we manually annotate 100 Chinese abstracts in the EECS domain as the ground-truth. The statistics of both datasets are given in Table 1. From UM-Corpus, a large-scale English-Chinese parallel corpus [Tian *et al.*, 2014], we select 30,000 English and Chinese sentences in the Thesis domain for training the language discriminator.

As shown in Table 2, we evaluate our method in different settings. The main network for information structure identification is denoted as `MAIN`, the main network that is co-trained with the language discriminator is denoted as `MAIN+ADV`, and `SEQ` denotes that sequence modeling on the top of the model is employed to maximize the probability of the whole abstract. The Viterbi algorithm is performed to find the most likely sequence. En/En and Ch/Ch denote the monolingual performances by 10-fold cross-validation on the English and the Chinese datasets, respectively. En/Ch denotes the cross-lingual performances, where the model is trained on the English data and tested on the Chinese data.

The cross-lingual `MAIN` model achieves an F-score of 58.17%, which is slightly superior to that of its monolingual counterpart trained on the small Chinese dataset. This result shows that the basic model, which regards the information from bilingual word embeddings and the Universal POS tags, is capable of cross-lingual generalization. Sequential modeling is useful for both monolingual and cross-lingual models.

[3] http://writcent.nthu.edu.tw/writcent/

Model	En/En	Ch/Ch	En/Ch
`MAIN`	65.60%	57.06%	58.17%
`MAIN+SEQ`	71.58%	62.09%	66.10%
`MAIN+ADV`	N/A	N/A	63.88%
`MAIN+ADV+SEQ`	N/A	N/A	71.93%

Table 2: Performances of our method in different settings. Marco-averaged F-scores are reported.

The language-adversarial learning strategy further improves the `MAIN` and `MAIN+SEQ` models. Co-training with a language discriminator successfully reduces the performance gap between the source and the target languages. The adversarial model with sequence labeling achieves a promising F-score of 71.93%, making a significantly improvement over all other models for the target language ($p < 0.01$).

4 Cross-lingual Knowledge Retrieval

We construct a KB with the results extracted by information structure identification. For each sentence in a collection of EECS abstracts written in both English and Chinese, we label its information structure type and insert it into the KB as a node. The relation between a node pair in the KB is denoted by the tuple of their types. For example, $\langle Purpose, Method \rangle$ denotes a relation between nodes a and b, where b is the method for the research goal a.

For an abstract submitted to CISA, each sentence will be labeled with its information structure type and linked to the related nodes with the same type in the KB. Various types of knowledge stored in the KB will be retrieved by using graph traversal. Finally, CISA will list the feasible approaches to user's research goal, the potential applications of the proposed method, and so on.

For the linking of cross-lingual information, the sentences in both English and Chinese are represented by using the bilingual word embeddings. Once all the nodes in the KB and the sentences in the submitted abstract are encoded as vectors in the same space, their relatedness can be computed by using cosine similarity. An evaluation on 100 abstracts shows our sentence representation achieves an MRR of 0.70 in cross-lingual sentence retrieval.

5 Conclusion

This work presents CISA, the first system for Chinese information structure analysis. We propose a model for Chinese information structure identification without the need of labeled Chinese data. The multi-label, finer-grained information structure analysis will be explored in the future.

Acknowledgements

This research was partially supported by Ministry of Science and Technology, Taiwan, under grants MOST-105-2221-E-002-154-MY3, MOST-106-2923-E-002-012-MY3 and MOST-107-2634-F-002-011-. We thank the Writing Center at National Tsing Hua University for providing us the corpus.

References

[Chen *et al.*, 2016] Xilun Chen, Ben Athiwaratkun, Yu Sun, Kilian Q. Weinberger, and Claire Cardie. Adversarial deep averaging networks for cross-lingual sentiment classification. *CoRR*, abs/1606.01614, 2016.

[Cho *et al.*, 2014] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics.

[Contractor *et al.*, 2012] Danish Contractor, Yufan Guo, and Anna Korhonen. Using argumentative zones for extractive summarization of scientific articles. In *Proceedings of COLING 2012*, pages 663–678, Mumbai, India, December 2012. The COLING 2012 Organizing Committee.

[Gouws *et al.*, 2015] Stephan Gouws, Yoshua Bengio, and Greg Corrado. Bilbowa: Fast bilingual distributed representations without word alignments. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 748–756, Lille, France, 07–09 Jul 2015. PMLR.

[Guo *et al.*, 2014] Yufan Guo, Diarmuid Ó Séaghdha, Ilona Silins, Lin Sun, Johan Högberg, Ulla Stenius, and Anna Korhonen. Crab 2.0: A text mining tool for supporting literature review in chemical cancer risk assessment. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, pages 76–80. Dublin City University and Association for Computational Linguistics, 2014.

[Huang and Chen, 2017] Hen-Hsen Huang and Hsin-Hsi Chen. Disa: A scientific writing advisor with deep information structure analysis. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 5229–5231, 2017.

[Joty *et al.*, 2017] Shafiq Joty, Preslav Nakov, Lluís Màrquez, and Israa Jaradat. Cross-language learning with adversarial neural networks. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 226–237, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[Kim *et al.*, 2017] Joo-Kyung Kim, Young-Bum Kim, Ruhi Sarikaya, and Eric Fosler-Lussier. Cross-lingual transfer learning for pos tagging without cross-lingual resources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2832–2838, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[Lee *et al.*, 2016] Lung-Hao Lee, Gaoqi RAO, Liang-Chih Yu, Endong XUN, Baolin Zhang, and Li-Ping Chang. Overview of nlp-tea 2016 shared task for chinese grammatical error diagnosis. In *Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA2016)*, pages 40–48, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee.

[Shiue *et al.*, 2017] Yow-Ting Shiue, Hen-Hsen Huang, and Hsin-Hsi Chen. Detection of chinese word usage errors for non-native chinese learners with bidirectional lstm. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 404–410. Association for Computational Linguistics, 2017.

[Teufel, 2006] Simone Teufel. *Argumentative Zoning for Improved Citation Indexing*, pages 159–169. Springer Netherlands, Dordrecht, 2006.

[Tian *et al.*, 2014] Liang Tian, Derek F. Wong, Lidia S. Chao, Paulo Quaresma, Francisco Oliveira, and Lu Yi. Um-corpus: A large english-chinese parallel corpus for statistical machine translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. European Language Resources Association (ELRA), 2014.

[Vulić and Moens, 2015] Ivan Vulić and Marie-Francine Moens. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 363–372, New York, NY, USA, 2015. ACM.