DISA: A Scientific Writing Advisor with Deep Information Structure Analysis (Demonstration)

Hen-Hsen Huang and **Hsin-Hsi Chen**
National Taiwan University, Taipei, Taiwan
hhhuang@nlg.csie.ntu.edu.tw; hhchen@ntu.edu.tw

Abstract

This paper demonstrates DISA, a higher-level writing assistant system, which analyzes the information structure of abstracts, and retrieves the knowledge according to the research goals from the related work. By incorporating the latest neural-network technologies including linguistically-informed neural-network and autoencoder, we construct an intelligent system which extends the scope of computer-aided writing.

1 Introduction

Most academic writing assistant systems aim to help users write better articles in terms of word usage and grammatical correctness [Chen et al., 2012; Dai et al., 2014; Liu et al., 2016]. This paper presents a writing assistant system that provides writing advice at a higher level. Rather than spelling checking and grammatical error diagnosis, our system, DISA[1] (deep information structure analysis), analyzes the information structure of a given article, and retrieves the useful knowledge in the related work for references.

In scientific writing, each sentence takes a different role to convey different information to readers [Swales, 1990]. Basic types of information structure such as Background, Purpose, Method, Results, and Conclusion are frequently used. Without a proper arrangement of information structure, a paper might be pointless even if it is flawless in terms of grammatical correctness.

Previous studies on information structure identification are applied to document summarization [Teufel, 2010; Contractor et al., 2012] and reading assistance [Guo et al., 2014]. In this work, we apply the structure information identification to writing assistance and show its effectiveness in knowledge retrieval.

The overview of DISA is illustrated in Figure 1. A web-based interface is provided for users to submit an abstract. The preprocessor first segments the abstract into sentences, and performs tokenization, part-of-speech tagging, and dependency parsing. Then, linguistic features like surface features and syntactic features are extracted for information structure identification. In the structure identification, the role of each sentence in the submitted abstract is labelled, and the research goals in the abstract can be found by locating the sentences which are labeled as Purpose. According to the research goals, our related work retrieval module ranks the most related passages from a collection of pre-analyzed EECS papers. In addition to the related works, feasible approaches to the research problem are also suggested. At the same time, the arrangement of the information structure in the abstract is also analyzed and compared to the norm.

Neural network-based approaches are investigated for information structure identification and knowledge retrieval. A linguistically-informed neural network model is adopted for information structure identification. For related word ranking and knowledge extraction, a GRU-based autoencoder is trained to model the similarity between sentences at the semantic level. Our final retrieval model is a hybrid of autoencoder and traditional information retrieval models. From the collection of pre-analyzed EECS papers, the retrieval model suggests the related works and the methods that are most related to the research goals in the submitted abstract.

The contributions of this work are three-fold. 1) We demonstrate a novel writing assistant system for higher level writing advising. 2) A new application of information structure identification is introduced to extend the scope of writing assistant systems. 3) Latest neural-network technologies are explored for performance improvement.

2 Resources

In the initial stage of the DISA project, we focus on the EECS area. NTHU Academic Writing Database[2], which contains 210,771 sentences from 4,380 EECS papers, is adopted as the

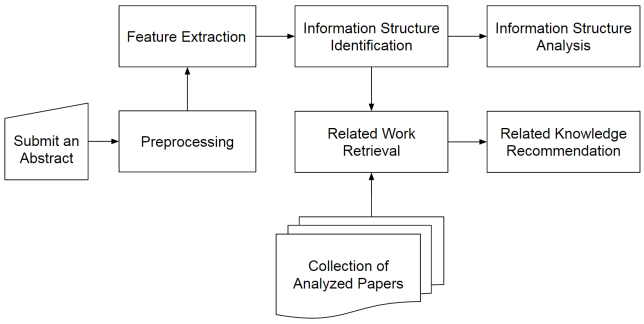

Figure 1: Flowchart of the DISA system.

[1] http://nlg18.csie.ntu.edu.tw/disa

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

Category	#	Description
Background	471	Description of the research issues; Contrast between the work and related work
Purpose	592	Research goals
Method	1,121	Description of data; Description of the experiments; Description of the data analysis
Results	1,025	Experimental findings; Discussion of results
Conclusion	185	Summary of the work; Future work

Table 1: Information structures annotated in the NTHU dataset.

research material. In this dataset, 3,394 sentences from 597 abstracts are annotated with the information structure labels in five categories by linguistic experts. Table 1 summarizes the categories. We train the information structure model with the labeled sentences, and train the autoencoder with all sentences in the dataset.

3 Information Structure Identification

Information structure identification is a task of sentence classification. In previous works, the computational models for information structure identification are mostly based on the feature-based learning approach. Various linguistic features and learning strategies are explored [Guo et al. 2011; Seaghdha and Teufel, 2014].

Linguistically informed neural-network models [Ebert et al., 2015], which are trained with the raw text and handcrafted linguistic features, benefit from taking both the raw data and the human-inspired information into account. We incorporate a linguistically informed model for information structure identification into DISA. The linguistic features suggested by Guo et al. [2013] are extracted, including sentence location, n-gram, part-of-speech tag, and dependency relation. The Stanford CoreNLP server is integrated for basic language processing [Manning et al., 2014]. Moreover, the Viterbi algorithm is performed on the top of the neural network for sequential modeling. For example, the sentences of Result are more likely to succeed the sentences of Method. We train the model on the NTHU dataset. The 10-fold cross validation shows the model achieves a macro-average f-score of 75.04% in five-way classification.

The abstracts with extremely poor structure arrangement may mislead the sequential model. For such cases, the sequential modeling can be disabled, and each sentence will be independently labeled.

4 Integrated System

Based on the outcome of information structure identification, two kinds of writing advices are made by the DISA system.

4.1 Advice on Structure Arrangement

The distributions of information structure in the dataset are calculated as the norm. Instead of the sentence frequency of each category provided in Table 1, we count the ratio of each category of information structure appearing in abstracts (i.e., document frequency). The document frequencies of Background, Purpose, Method, Results, and Conclusion are 33.17%, 93.63%, 70.85%, 80.07%, and 23.28%, respectively. The statistics shows that an abstract without a statement of Purpose or Results may be problematic.

Once the information structure of a submitted abstract is identified, the distribution of information structure in the abstract will be compared to the norm by using Euclidean distance. For an abstract with an abnormal structure arrangement, our system will generate an advice. For example, DISA will remind the user to add a sentence of Purpose in an abstract lacking of Purpose.

4.2 Knowledge Retrieval

The traditional information retrieval model represents the data by using the "bag-of-words" scheme, where the word order information is disregarded. By contrast, the sequential autoencoder, which projects each piece of data into a dense vector space, can better utilize the information in the sequential data like sentences [Dai and Le, 2015].

In this work, we train a GRU [Cho et al., 2014] autoencoder that represents a sentence as a 50-dimensional vector. With additional features including TF-IDF weights, each sentence in the database is encoded as a vector. The sentences in the submitted abstract are also encoded to vectors in the same vector space, and then the distance between a user-submitted sentence and a sentence in the database can be measured using cosine similarity. Based on the type of information structures, the related Methods for the Purpose in the submitted abstract, the related Backgrounds for the Purpose in the submitted abstract, and the Purposes similar to the one in the submitted abstract are retrieved.

From the retrieved sentences, we further apply the topic model for result filtering at the concept level. The LDA [Blei et al., 2003] model is learned to cluster the abstracts from the NTHU dataset into topics like "data mining", "signal processing", and "photonics". The DISA system infers the topic of the user-submitted abstract by using the LDA model, and truncates the retrieval results which are not on the same topic.

5 Conclusion

We present an innovative computer-aided writing system aiming at providing advice at the structure and the knowledge levels. Rather than aiding users to compose right sentences, our system assists users to create good research work.

As a novel system in its initial stage, the information structure detector is trained on a moderate-sized EECS dataset, and the capability of knowledge retrieval is also limited by the size and the diversity of the backend database. In the future, more scientific articles from the public domain will be incrementally imported to the database. User study will be conducted to measure the performance of our system in the real world.

Acknowledgements

This research was partially supported by Ministry of Science and Technology, Taiwan, under grants MOST-104-2221-E-002-061-MY3 and MOST-105-2221-E-002-154-MY3. We thank the Writing Center at National Tsing Hua University for providing us the NTHU Academic Writing Database.

References

[Blei et al., 2003] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research 3*:993-1022, 2003.

[Chen et al., 2012] Mei-Hua Chen, Shih-Ting Huang, Hung-Ting Hsieh, Ting-Hui Kao, Jason S. Chang. FLOW: A First-Language-Oriented Writing Assistant System. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics* (*ACL 2012*): *System Demonstrations*, pages 157–162, Jeju, Republic of Korea, 2012.

[Cho et al., 2014] Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint, arXiv*:1409.1259, 2014.

[Contractor et al., 2012] Danish Contractor, Yufan Guo, and Anna Korhonen. Using Argumentative Zones for Extractive Summarization of Scientific Articles. In *Proceedings of the 24th International Conference on Computational Linguistics* (*COLING 2012*): *Technical Papers*, pages 663–678, Mumbai, India, 2012.

[Dai et al., 2014] Xianjun Dai, Yuanchao Liu, Xiaolong Wang, and Bingquan Liu. WINGS: Writing with Intelligent Guidance and Suggestions. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics* (*ACL 2014*): *System Demonstrations*, pages 25–30, Baltimore, Maryland USA, 2014.

[Ebert et al., 2015] Sebastian Ebert, Ngoc Thang Vu, and Hinrich Schütze. A Linguistically Informed Convolutional Neural Network. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (*WASSA 2015*), pages 109–114, Lisboa, Portugal, 2015.

[Guo et al., 2011] Yufan Guo, Anna Korhonen, and Thierry Poibeau. A Weakly-supervised Approach to Argumentative Zoning of Scientific Documents. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (*EMNLP 2011*), pages 273–283, Edinburgh, Scotland, UK, 2011.

[Guo et al., 2013] Yufan Guo, Roi Reichart, and Anna Korhonen. Improved Information Structure Analysis of Scientific Documents Through Discourse and Lexical Constraints. In *Proceedings of NAACL-HLT 2013*, pages 928–937, Atlanta, Georgia, 2013.

[Guo et al., 2014] Yufan Guo, Diarmuid O Seaghdha, Ilona Silins, Lin Sun, Johan Hogberg, Ulla Stenius, and Anna Korhonen. CRAB 2.0: A text mining tool for supporting literature review in chemical cancer risk assessment. In *Proceedings of the 25th International Conference on Computational Linguistics* (*COLING 2014*): *System Demonstrations*, pages 76–80, Dublin, Ireland, 2014.

[Liu et al., 2016] Yuanchao Liu, Xin Wang, Ming Liu, Xiaolong Wang. Write-righter: An Academic Writing Assistant System. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence* (*AAAI 2016*), pages 4373-4374, 2016.

[Manning et al., 2014] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (*ACL 2014*): *System Demonstrations*, pages 55-60, Baltimore, Maryland USA, 2014.

[Seaghdha and Teufel, 2014] Diarmuid O Seaghdha and Simone Teufel. Unsupervised learning of rhetorical structure with un-topic models. In *Proceedings of the 25th International Conference on Computational Linguistics* (*COLING 2014*): *Technical Papers*, pages 2–13, Dublin, Ireland, 2014.

[Swales, 1990] John M. Swales. *Genre analysis: English in academic and research settings*. Cambridge University Press, Cambridge, UK, 1990.

[Teufel, 2010] Simone Teufel. 2010. *The Structure of Scientific Articles: Applications to Citation Indexing and Summarization*. CSLI Publications, Stanford, CA, 2010.