RAAMove: A Corpus for Analyzing Moves in Research Article Abstracts

Hongzheng Li[*] Ruojin Wang[*], Ge Shi[†], Xing Lv[*], Lei Lei[*], Chong Feng[‡], Fang Liu[*], Jinkun Lin[‡], Yangguang Mei[†], Lingnan Xu[‡]

[*]School of Foreign Languages,
Key Laboratory of Language, Cognition and Computation Ministry of Industry and Information Technology,
Beijing Institute of Technology
[†]Faculty of Information Technology, Beijing University of Technology
[‡]School of Computer Science, Beijing Institute of Technology
[*]8, Liangxiang East Road, Fangshan District, 102488, Beijing
[‡]5, Zhongguancun South St., Haidian District, 100081, Beijing, China
[†]100, Pingleyuan, Chaoyang District, 100124, Beijing, China
{lihongzheng,wangruojin,lvxing,leilei,fengchong,liufang,ljkun,xln}@bit.edu.cn
{shige,meiyangguang}@bjut.edu.cn

Contents:

1. Introduction

Background:

(1) importance of research articles (RAs)

(2) move analysis for instructing, assessing abstract writing

Gap:

(1) few move analysis targeting multi-disciplinary RAs

(2) few large-scale annotated corpora for RA abstracts

[**Background**]Large-scale pretrained language models have achieved SOTA results on NLP tasks. [**Gap**]However, they have been shown vulnerable to adversarial attacks especially for logographic languages like Chinese. [**Purpose**]In this work, we propose RoCBert: a pretrained Chinese Bert that is robust to various forms of adversarial attacks like word perturbation, synonyms, typos, etc. [**Method**]It is pretrained with the contrastive learning objective which maximizes the label consistency under different synthesized adversarial examples. ...... [**Result**]Across 5 Chinese NLU tasks, RoCBert outperforms strong baselines under three blackbox adversarial algorithms without sacrificing the performance on clean testset.

1. Introduction

- We develop a multi-domain move structure annotation corpus for analyzing moves in RA abstracts.
- We suggest a revision of move structure categories based on Hyland's established move classification (Hyland, 2000).
- We propose an innovative BERT-based automatic annotation model that incorporates word level saliency attribution.

2. Scheme

This section attempts to answer the following questions:

(1) Which rhetorical move theories could guide the construction of a corpus for analyzing moves in RA abstracts?

(2) In what manner can these theories find greater relevance in our work?

2. Scheme

(1) Which rhetorical move theories could guide the construction of a corpus for analyzing moves in RA abstracts?

Move	Function
Introduction	Establishes context of the paper and motivates the research or discussion.
Purpose	Indicates purpose, thesis or hypothesis, outlines the intention behind the paper.
Method	Provides information on design, procedures, assumptions, approach, data, etc.
Product	States main findings or results, the argument, or what was accomplished.
Conclusion	Interprets or extends results beyond scope of paper, draws inferences, points to applications or wider implications.

Table 1: Hyland's classification of rhetorical moves in RA abstracts

2. Scheme

(1) Which rhetorical move theories could guide the construction of a corpus for analyzing moves in RA abstracts?

	AI		Engineering	
Move	**Freq.**	**%**	**Freq.**	**%**
Intro.	17	85%	16	80%
Pur.	20	100%	19	95%
Met.	19	95%	19	95%
Pro.	11	55%	16	80%
Con.	11	55%	13	65%

Table 2: Frequency of moves identified based on Hyland's classification in sample abstracts. Where the five moves correspond to those in Table 1.

[Example 1]: These results can provide a better understanding of surfactants and guide the practical preparation of multicomponent fluids for boiling heat transfer enhancement.

[Example 2]: We release source code for our models and experiments at https://github.com/xxx.

[Example 3]: Undermining the impact of hateful content with informed and non-aggressive responses, called counter-narratives, has emerged as a possible solution for having healthier online communities.

[Example 4]: Although such studies have made an effort to build hate speech / counternarrative (HS/CN) datasets for neural generation, they fall short in reaching either high-quality and/or high-quantity.

2. Scheme

(2) In what manner can these theories find greater relevance in our work?

Move	Function
Background	States the research area and provides any historical, theoretical, or empirical related information.
Gap	Establishes a niche: indicates a gap, adds to what is known, presents positive justification (Swales, 2004).
Purpose	Indicates purpose, thesis or hypothesis, outlines the intention behind the paper.
Method	Provides information on design, procedures, assumptions, approach, data, etc.
Result	States main findings or results or what was accomplished.
Conclusion	Summarizes the results or extends results beyond scope of paper.
Implication	Draws inferences which has not been explicitly stated.
Contribution	Points out the theoretical and practical value.

Table 3: Enriched move classification

3. Corpus Construction

Two phases:

(1) data selection and preprocessing

(2) process (manual annotation + automatic annotation)

3. Corpus Construction

(1) data selection and preprocessing

Discipline	Journal/Conference
Artificial Intelligence	the Annual Meeting of the Association for Computational Linguistics (ACL)
Artificial Intelligence	Technical Track on CV on the AAAI Conference on Artificial Intelligence (AAAI)
Mechanical Engineering	Journal of Mechanical Design
Mechanical Engineering	International Journal of Heat and Mass Transfer
Communication Engineering	IEEE Journal on Selected Areas in Communications

Table 4: Selected journals and conferences for annotation

3. Corpus Construction

(2) process: manual annotation

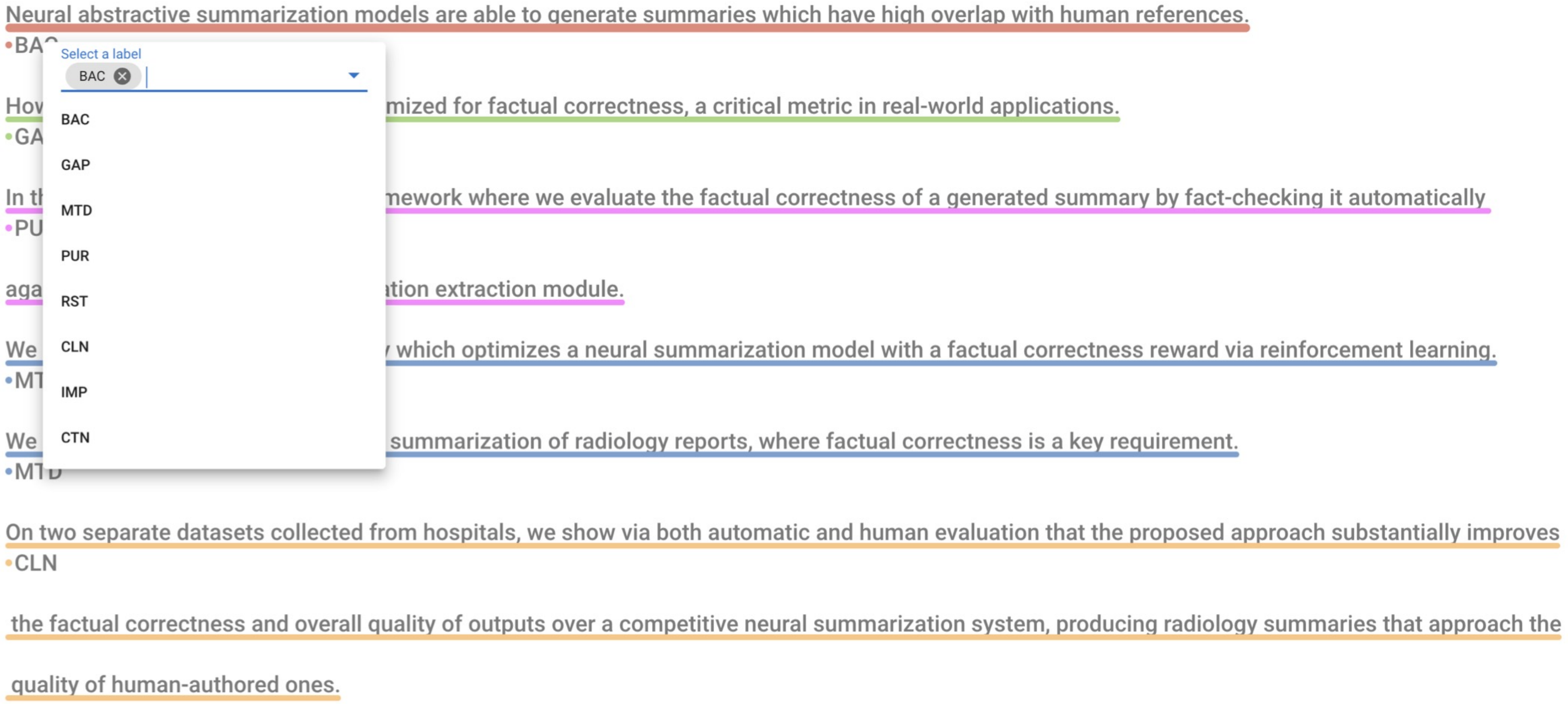

Figure 1: Screenshot of the doccano annotation platform

https://github.com/doccano/doccano

3. Corpus Construction

(2) process (manual annotation + automatic annotation): guidelines

Label	Abbreviation
Background	BAC
Gap	GAP
Purpose	PUR
Method	MTD
Result	RST
Conclusion	CLN
Implication	IMP
Contribution	CTN

Table 5: Annotation labels and their abbreviation

[**BAC**]While neural networks with attention mechanisms have achieved superior performance on many natural language processing tasks, [**GAP**]it remains unclear to which extent learned attention resembles human visual attention.

3. Corpus Construction

(2) process: manual annotation

expert team + weekly discussions + revision

3. Corpus Construction

(2) process:

automatic annotation: BERT-based model

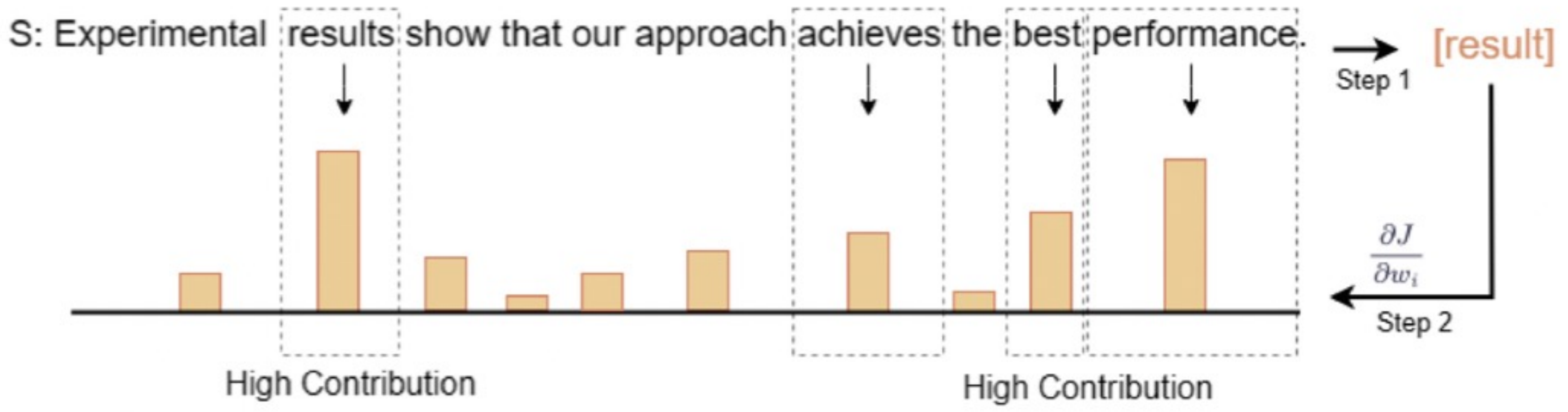

Figure 2: An illustration of move saliency attribution

4. Corpus Statistics

(1) distribution of move types

Label	Frequency	%
BAC	6,466	19.02
GAP	3,272	9.63
PUR	4,874	14.34
MTD	**11,526**	**33.91**
RST	3,732	10.98
CLN	3,006	8.84
IMP	282	0.83
CTN	830	2.44
Total	33,988	100

Table 6: Frequency and distribution of moves identified in our corpus

(2) occurrence of move types

	AI		Engineering	
Move	**#**	**%**	**#**	**%**
BAC	2,003	75.02	1,528	76.29
GAP	1,518	56.85	891	44.48
PUR	**2,333**	**87.38**	**1,901**	**94.91**
MTD	2,245	84.08	1,873	93.51
RST	1,540	57.68	953	47.58
CLN	1,192	44.64	1,079	53.87
IMP	112	4.19	159	7.94
CTN	544	20.37	215	10.73

Table 7: Occurrence and distribution of each move type identified across the two fields in our corpus

4. Corpus Statistics

	AI	Engineering
#Sent.	17,391	16,597
Average #Sent.	6.51	8.29
#Words	381,734	406,244
Average #Words	142.97	202.82
Average #Move types	4.38	4.29

Table 8: The average number of sentences, words, and move types in each abstract within the two fields

5. Experiments

(1) move recognition

Data	#Sentences
Training set	7,147
Test set	1,787

Table 9: Dataset statistics

Method	P (%)	R (%)	F1 (%)
BERT	74.06	79.58	76.72
BERT+Context	74.55	81.23	77.60
Our	75.01	82.34	**78.53**

Table 10: Results of move structure identification

(2) comparison with ChatGPT

BO

The move structure of a scientific paper refers to the categorical composition of the linguistic rhetorical components of the academic discourse in the paper. Move recognition is essentially a classification problem in sentences. Now the moves are background, gap, method, purpose, result, conclusion, contribution, and implication. Here are a few examples of move recognition: Detecting emotion in text allows social and computational scientists to study how people behave and react to online events. [background] However, developing these tools for different languages requires data that is not always available. [gap] This paper collects the available motion detection datasets across 19 languages. [purpose] We train a multilingual emotion prediction model for social media data, XLM-EMO. [method] The model shows competitive performance in a zero-shot setting, suggesting it is helpful in the context of low-resource languages. [results] We release our model to the community so that interested researchers can directly use it. [contribution] Below I will give you some sentences, these sentences are from scientific papers, please complete the step recognition.

“Input sentences”

What is the move of this sentence?

Figure 3: Instructions for ChatGPT

Our VS. ChatGPT

80 VS. 65

Thank you!