Overview of CCL24-Eval Task6: Chinese Essay Rhetoric Recognition and Understanding (CERRU)

Nuowei Liu[1], Xinhao Chen[1], Yupei Ren[1,2], Man Lan[1,2]*,
Xiaopeng Bai[2,3], Yuanbin Wu[1,2], Shaoguang Mao[4], Yan Xia[4]
[1]School of Computer Science and Technology, East China Normal University
[2]Shanghai Institute of AI for Education, East China Normal University
[3]Department of Chinese Language and Literature, East China Normal University
[4]Microsoft Research Asia
{nwliu, 51215901006, ypren}@stu.ecnu.edu.cn
{mlan, ybwu}@cs.ecnu.edu.cn, xpbai@zhwx.ecnu.edu.cn
{shaoguang.mao, yanxia}@microsoft.com

Abstract

Rhetoric is fundamental to the reading comprehension and writing skills of primary and middle school students. However, current work independently recognize single coarse-grained categories or fine-grained categories. In this paper, we propose the CCL24-Eval Task6: Chinese Essay Rhetoric Recognition and Understanding (CERRU), consisting of 3 tracks: (1) Fine-grained Form-level Categories Recognition, (2) Fine-grained Content-level Categories Recognition and (3) Rhetorical Component Extraction. A total of 32 teams registered to participate in CERRU and 9 teams submitted evaluation results, with 7 of these teams achieving an overall score that surpassed the baseline.

1 Introduction

In the learning process of primary and middle school students, rhetoric is not only a core component of reading comprehension and writing skills, but also an indispensable element in shaping excellent literary works. Recognizing and understanding the use of rhetoric in students' essays can help improve their expressive abilities in writing. However, this requires a significant amount of manual effort, posing challenges to teachers in term of essay assessment and instruction. With the development of education and the widespread availability of the Internet, many researchers have begun to explore the use of computer technology for automatic grading of essays (Rudner et al., 2006), where the use of rhetoric is a crucial part of teachers' essay grading.

The use of rhetoric in essays reflects the level of literacy grace and language expression ability (Guo et al., 2018), which is significant for helping teachers assess the quality of essays and guide students in improving their expressive skills. In recent years, research on the recognition of rhetoric in essays often employs alignment strategies and other rules to perform coarse-grained recognition of rhetoric such as parallelism and metaphor from the perspectives of sentence structure and semantic information (Niculae, 2013; Song et al., 2016) or designs model structures specifically to recognize simile (Liu et al., 2018; Zeng et al., 2020). These efforts independently recognize different major rhetorical categories such as metaphor, personification, hyperbole and parallelism, lacking universality. On the other hand, they are coarse-grained and lack fine-grained definitions of rhetorical categories. Furthermore, beyond recognizing rhetorical categories in sentences, some researches treat the understanding of rhetoric as a component extraction task, for example, extracting the tenor and the vehicle from metaphorical sentences (Wang et al., 2022). These researches lack definitions for the rhetorical subjects and contents of other rhetorical devices, and thus cannot provide systematic and comprehensive guidance and feedback on the essays of elementary and middle school students.

Therefore, to address the aforementioned challenges, we propose the CCL24-Eval Task6: **C**hinese **E**ssay **R**hetoric **R**ecognition and **U**nderstanding (CERRU). The dataset for the evaluation originates from

*Corresponding author
©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

examination essays written by elementary and middle school students whose native language is Chinese. The genres of these essays include narrative and argumentative writing, among others. Our task settings systematically define the fine-grained rhetorical categories found in these essays, recognizing them from both form level and content level based on the linguistic definitions of rhetoric (Li, 2020). Furthermore, we define the subjects and contents of each rhetorical category, which aids teachers in understanding the use of rhetoric at the sentence level in student essays. It also supports elementary and middle school students in practicing appropriate rhetorical techniques in their writing.

CERRU categorizes rhetorical devices into metaphor, personification, hyperbole and parallelism, and further subdivides these four rhetorical categories into fine-grained categories. As shown in Figure 1, CERRU includes 3 tracks, which are

- Track1: Fine-grained Form-level Categories Recognition
- Track2: Fine-grained Content-level Categories Recognition
- Track3: Rhetorical Component Extraction

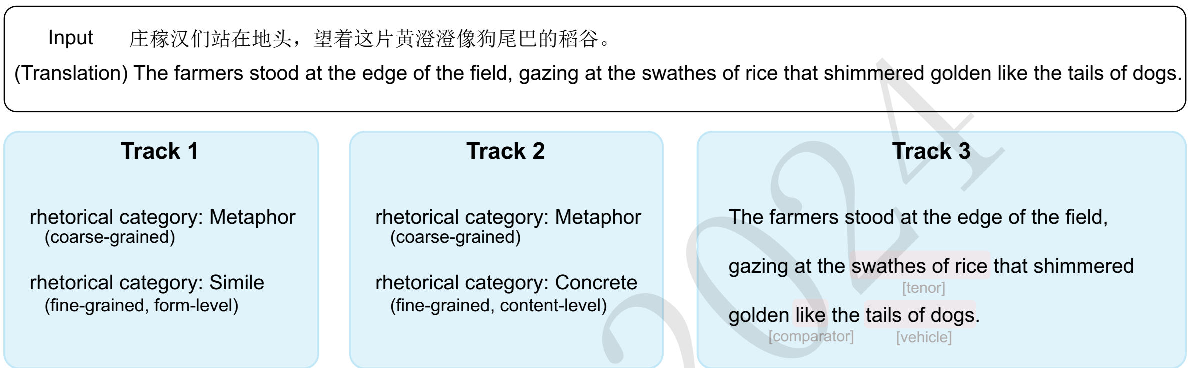

Figure 1: An example of CERRU.

CCL 2024

2 Task Descriptions

2.1 Track1: Fine-grained Form-level Categories Recognition

Track1 uses sentences as basic units and categorizes the rhetorical devices into four coarse-grained categories: metaphor, personification, hyperbole, parallelism. As shown in Table 1, each category is further subdivided into fine-grained form-level categories.

- For metaphor, it is subdivided into simile, metaphor and metonymy.
- For personification, it is subdivided into noun, verb, adjective and adverb.
- For hyperbole, it is subdivided into direct hyperbole, indirect hyperbole and mixed hyperbole.
- For parallelism, it is subdivided into structure parallelism and sentence parallelism.

Metaphor			Personification				Hyperbole			Parallelism	
Simile	Metaphor	Metonymy	Noun	Verb	Adjective	Adverb	Direct Hyperbole	Indirect Hyperbole	Mixed Hyperbole	Structure Parallelism	Sentence Parallelism

Table 1: The relationship between coarse-grained categories and fine-grained form-level categories.

Track1 is a multi-label classification problem, involving predicting the coarse-grained rhetorical category and fine-grained form-level category used in a given sentence.

2.2 Track2: Fine-grained Content-level Categories Recognition

Similar to track1, track2 uses sentences as basic units and categorizes the rhetorical devices into four coarse-grained categories: metaphor, personification, hyperbole, parallelism. As shown in Table 2, each category is further subdivided into fine-grained content-level categories.

- For metaphor, it is subdivided into concrete, action and abstract.
- For personification, it is subdivided into personification and anthropomorphism.
- For hyperbole, it is subdivided into amplification, understatement and prolepsis.
- For parallelism, it is subdivided into coordination, subordination and gradation.

Metaphor			Personification		Hyperbole			Parallelism		
Concrete	Action	Abstract	Personification	Anthro-pomorphism	Ampli-fication	Understatement	Prolepsis	Coor-dination	Subor-dination	Gradation

Table 2: The relationship between coarse-grained categories and fine-grained content-level categories.

Track2 is a multi-label classification problem, involving predicting the coarse-grained rhetorical category and fine-grained content-level category used in a given sentence.

2.3 Track3: Rhetorical Component Extraction

Rhetorical components include the described object in the given sentence and the specific content of the description. Extracting these components helps understanding students' use of rhetoric, reflecting their language expression skills. As shown in Table 3, track3 uses sentences as basic units and categorizes the rhetorical components in the sentences into connector, object and content.

- For metaphor-simile, the rhetorical components include comparator, tenor and vehicle. For metaphor-metaphor, the rhetorical components include tenor and vehicle. For metaphor-metonymy, the rhetorical components include vehicle.
- For personification, regardless of form-level category, the rhetorical components include personification object and personification content.
- For hyperbole, regardless of form-level category, the rhetorical components include hyperbole object and hyperbole content.
- For parallelism, regardless of form-level category, the rhetorical components include parallelism marker.

Rhetorical Component	Metaphor			Personification	Hyperbole	Parallelism
	Simile	Metaphor	Metonymy			
Connector	Comparator	-	-	-	-	Parallelism Marker
Object	Tenor	Tenor	-	Personification Object	Hyperbole Object	-
Content	Vehicle	Vehicle	Vehicle	Personification Content	Hyperbole Content	-

Table 3: Rhetorical components of different fine-grained form-level categories.

CCL 2024

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 253-261, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

3 Datasets

3.1 Dataset Annotation

CERRU collects the eesays used in our dataset from essays written by primary and middle school students for their exams. The collected data covers various geners of writing, such as character and scene description.

During the process of dataset annotation, four annotators participated, including undergraduates and postgraduates majoring in linguistics. First, preliminary annotation guidelines were established. Second, the four annotators jointly pre-annotated 50 essays. After completing the pre-annotation, the inter-annotator agreement of the annotation results was checked, and the annotation guidelines were further revised based on the results. Finally, each of the four annotators formally annotated about 140 essays, totaling 503 essays. Specifically, the last 20 essays annotated by Annotator A were identical to the first 20 essays annotated by Annotator B, and so on. The overlapped annotations were used to check the inter-annotator agreement of the formal annotation results.

3.2 Dataset Statistics

Track1, track2 and track3 share the same training set, validation set and test set while each track has distinct annotations. Track1 and track2 focus on fine-grained form-level and content-level categories respectively while track3 focus on rhetorical components. The size of dataset in shown in Table 4 and the portion of the test set used for evaluation constitutes approximately 10% of the entire test set.

#Training set	#Validation set	#Test set
634	225	5000

Table 4: Statistics of dataset used in CERRU.

4 Evaluation Metrics

In this section, we introduce the metrics used in CERRU. F_1 refers to macro-F1 score in track1, track2 and track3. The overall score of CERRU is the arithmetic mean of track1, track2 and track3.

4.1 Track1: Fine-grained Form-level Categories Recognition

As displayed in Equation 1, the overall F1 score of track1 is comprised of two parts: the F1 score of coarse-grained categories and fine-grained form-level categories.

$$F_1 = 0.3 \times F_1^{\text{rhetorical}} + 0.7 \times F_1^{\text{form}} \tag{1}$$

where $F_1^{\text{rhetorical}}$ denotes the F1 score of coarse-grained categories and F_1^{form} denotes the F1 score of fine-grained form-level categories.

4.2 Track2: Fine-grained Content-level Categories Recognition

As displayed in Equation 2, the overall F1 score of track2 is comprised of two parts: the F1 score of coarse-grained categories and fine-grained content-level categories.

$$F_1 = 0.3 \times F_1^{\text{rhetorical}} + 0.7 \times F_1^{\text{content}} \tag{2}$$

where $F_1^{\text{rhetorical}}$ denotes the F1 score of coarse-grained categories and F_1^{content} denotes the F1 score of fine-grained content-level categories.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 253-261, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

CCL 2024

4.3 Track3: Rhetorical Component Extraction

As displayed in Equation 3, the overall F1 score of track3 is comprised of three parts: the F1 score of connectors, the F1 score of objects and the F1 score of contents.

$$F_1 = \frac{1}{3} \times F_1^{\text{connector}} + \frac{1}{3} \times F_1^{\text{object}} + \frac{1}{3} \times F_1^{\text{content}} \tag{3}$$

where $F_1^{\text{connector}}$, F_1^{object} and F_1^{content} denotes the F1 score of connectors, objects and contents respectively.

5 Baselines

In this section, we introduce the baseline approaches used in CERRU and the scores on track1, track2 and track3.

For track1 and track2, we take both the tasks as multi-label classification problems and fine-tune RoBERTa [1] (Liu et al., 2019) on the training set. A Dropout (Srivastava et al., 2014) layer and a linear layer are concatenated to RoBERTa, and the output after applying sigmoid function is used to represent the probabilities of each category in the given sentence. For track3, we take the task as named entity recognition and fine-tune RoBERTa on the training set. A Dropout layer and a linear layer are concatenated to RoBERTa. Furthermore, we utilize the IOB tagging format (Ramshaw and Marcus, 1999) to tag the comparator, tenor, vehicle, personification object, personification content, hyperbole object, hyperbole content and parallelism marker. The output from RoBERTa after applying argmax function is represented as an entity tag on each token. Subsequently, the consecutive "`B-`" prefix tag and "`I-`" prefix tag are combined to represent the corresponding rhetorical components.

As shown in Table 5, we report the baseline scores on both the validation set and the test set for reference.

Track	F1 (on validation set) (%)	F1 (on test set) (%)
Track1	38.11	45.66
Track2	35.28	56.89
Track3	21.29	20.85

Table 5: Baseline results on the validation set and the test set.

6 Results

In this section, we first discuss the overall results, including the statistics of the participating teams and their scores on each track (See Section 6.1). Considering the correlation between different tracks, most of the teams choose to combine the dataset from different tracks for joint training. Therefore, we then discuss the approaches they use respectively (See Section 6.2 - Section 6.6). Finally, an overall analysis will be discussed in Section 6.7.

6.1 Overall Results

For CCL24-Eval Task6, a total of 32 teams registered to participate in CERRU. Utimately, 9 teams submitted evaluation results and obtained valid scores, with 7 of these teams achieving an overall score that surpassed the baseline. Details are listed in Table 6.

Furthermore, the statistics on the usage of LLMs, external data and data augmentation methods by the top 5 teams based on their overall scores are listed in Table 7.

[1] https://huggingface.co/uer/chinese_roberta_L-12_H-768

CCL 2024

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 253-261, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Team Name	Track1 (%)	Track2 (%)	Track3 (%)	Score (%)
Zhengzhou University (ZZU)	61.30	62.29	75.28	66.29
Beijing Language and Culture University (BLCU)	59.20	60.92	77.96	66.03
iHuman Inc.	53.77	60.15	68.26	60.72
Central China Normal University (CCNU)	50.86	55.81	73.75	60.14
Zhongyuan University of Technology (ZUT1)	51.48	55.11	69.51	58.70
Zhongyuan University of Technology (ZUT2)	51.48	55.82	57.00	54.77
Institute of Computing Technology (ICT)	50.23	52.78	54.22	52.41
baseline	**45.66**	**56.89**	**20.85**	**41.13**
Individual Team	40.00	52.66	-	37.84
Jiangxi Normal University (JXNU)	39.60	39.13	-	33.19

Table 6: Scores of the participating teams. ”-” indicates that the team did not submit evaluation results on the track, and the overall score is calculated based on the baseline.

Team Name	LLMs	External Data	Data Augmentation
ZZU	✓	✗	✓
BLCU	✓	✓	✗
iHuman Inc.	✓	✗	✗
CCNU	✗	✗	✗
ZUT1	✓	✗	✓

Table 7: Statistics on the usage of LLMs, external data and data augmentation methods. ”LLMs” indicates whether to use Large Language Models. ”External Data” indicates whether data outside the provided dataset for CERRU is used. ”Data Augmentation” indicates whether any augmentation is performed on the provided dataset for CERRU.

CCL 2024

6.2 Team ZZU

ZZU employ LoRA (Hu et al., 2021) method for instruction fine-tuning Yi (Young et al., 2024) and Qwen1.5 (Team, 2024). Noticing that the three tracks share the same training set, validation set and test set, differing only in the respective annotations, they combine the instruction datasets from the three tracks and perform multi-task fine-tuning on the mixed dataset. Moreover, inspired by the LLM2LLM method (Lee et al., 2024), they record error-prone samples in track1 and track2 from the validation set during the fine-tuning process, using a more powerful LLM as a teacher model to generate synthetic data based on these error-prone samples. Additionally, to further enhance model performance, they explore a model ensemble approach to classify coarse-grained and fine-grained categories using LLMs.

6.3 Team BLCU

To expand the dataset, BLCU first adopt GLGC (A Corpus for Chinese Literary Grace Evaluation) (Li et al., 2022), a publicly available corpus, and some online data as the external data. Then, they propose an approach for Chinese rhetoric recognition and understanding with collaborative decision-making between large and small language models under the guidance of human thinking. They redefine the order of tasks and select the large and small language models in the specific process to reach the local optimization at each step. In particular, they use BERT (Devlin et al., 2018) to output the probabilities of each category in a given sentence and employ GPT-4 (Achiam et al., 2023) to predict the result using the output after applying the softmax function.

6.4 Team iHuman Inc.

iHuman Inc. directly employ LoRA (Hu et al., 2021) for fine-tuning Qwen-7B (Bai et al., 2023). For track1 and track2, they first predict the coarse-grained categories of each given sentence and then predict

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 253-261, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

the corresponding fine-grained categories of the given sentence. To enhance the robustness of their approach, multiple prompts are pre-defined. For track3, noticing that the predicted output may not be exactly the same as in the given sentence, they use a substring comparison method based on edit distance. Particularly, when the edit distance between the output and a substring of the input sentence is less than a certain threshold, they consider them to be identical and directly use the corresponding substring as the result.

6.5 Team CCNU

CCNU employ the unified multi-task learning architecture to fully incorporate the correlation between the three tracks. First, they use the Transformer (Vaswani et al., 2017) pre-trained model as shared feature encoder to represent the sentences. The framework they propose consists of four sub-tasks: rhetorical device recognition, form-level category recognition, content-level category recognition and rhetorical component extraction, which enhance each other's fusion learning. Finally, the aforementioned sub-tasks are integrated into a unified model through parameter sharing.

6.6 Team ZUT1

ZUT1 employ an ensemble model combining BERT (Devlin et al., 2018) and ERNIE (Sun et al., 2019) for track1 and track2. Furthermore, a data augmentation approach is used to enable the model to learn more relevant features from the imbalanced dataset. In particular, they apply methods such as synonym replacement, random word insertion and similar sentence generation to the labeled data. Additionally, they add the prediction generated by the model on unlabeled data back into the training set, thereby increasing the size of training set to enhance the performance of the model. For track3, they use ChatGLM-6B (Zeng et al., 2022) and Qwen-7B (Bai et al., 2023) with QLoRA (Dettmers et al., 2024) fine-tuning method to extract the rhetorical components from the given sentence.

6.7 Overall Analysis

Overall, the teams using LLMs perform better on most tracks compared to those using other approaches while CCNU also achieve a competitive performance. Additionally, the use of external data and data augmentation methods also significantly improves the performance. Most of the teams use LoRA or QLoRA to fine-tune the LLMs, while the methods of data augmentation vary between the teams. Furthermore, several teams improve the overall performance by effectively defining new sub-tasks and rearrange the order in which these sub-tasks are addressed.

7 Conclusion

In this paper, we propose the CCL24-Eval Task6: **C**hinese **E**ssay **R**hetoric **R**ecognition and **U**nderstanding (CERRU), consisting of 3 tracks: (1) Fine-grained Form-level Categories Recognition, (2) Fine-grained Content-level Categories Recognition and (3) Rhetorical Component Extraction. A total of 32 teams registered to participate in CERRU and 9 teams submitted evaluation results and obtained valid scores. Furthermore, we discuss the approaches used by the top 5 teams based on their overall scores. The results demonstrate that the usage of LLMs and data augmentation methods help improve the overall scores.

Acknowledgements

We appreciate the support from National Natural Science Foundation of China with the Main Research Project on Machine Behavior and Human Machine Collaborated Decision Making Methodology (72192820 & 72192824), Pudong New Area Science Technology Development Fund (PKX2021-R05), Science and Technology Commission of Shanghai Municipality (22DZ2229004) and Shanghai Trusted Industry Internet Software Collaborative Innovation Center.

CCL 2024

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jingjin Guo, Wei Song, Xianjun Liu, Lizhen Liu, and Xinlei Zhao. 2018. Attention-based bilstm network for chinese simile recognition. In *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)*, pages 144–147. IEEE.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Nicholas Lee, Thanakul Wattanawong, Sehoon Kim, Karttikeya Mangalam, Sheng Shen, Gopala Anumanchipali, Michael W Mahoney, Kurt Keutzer, and Amir Gholami. 2024. Llm2llm: Boosting llms with novel iterative data enhancement. *arXiv preprint arXiv:2403.15042*.

Yi Li, Dong Yu, and Pengyuan Liu. 2022. Clgc: A corpus for chinese literary grace evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5548–5556.

Qingrong Li. 2020. *Modern Practical Chinese Rhetoric*. BEIJING BOOK CO. INC. In Chinese.

Lizhen Liu, Xiao Hu, Wei Song, Ruiji Fu, Ting Liu, and Guoping Hu. 2018. Neural multitask learning for simile recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1543–1553.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Vlad Niculae. 2013. Comparison pattern matching and creative simile recognition. In *Proceedings of the Joint Symposium on Semantic Processing. Textual Inference and Structures in Corpora*, pages 110–114.

Lance A Ramshaw and Mitchell P Marcus. 1999. Text chunking using transformation-based learning. In *Natural language processing using very large corpora*, pages 157–176. Springer.

Lawrence M Rudner, Veronica Garcia, and Catherine Welch. 2006. An evaluation of intellimetric™ essay scoring system. *The Journal of Technology, Learning and Assessment*, 4(4).

Wei Song, Tong Liu, Ruiji Fu, Lizhen Liu, Hanshi Wang, and Ting Liu. 2016. Learning to identify sentence parallelism in student essays. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers*, pages 794–803.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*.

Qwen Team. 2024. Introducing qwen1.5, February.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Xiaoyue Wang, Linfeng Song, Xin Liu, Chulun Zhou, and Jinsong Su. 2022. Getting the most out of simile recognition. *arXiv preprint arXiv:2211.05984*.

CCL 2024

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Jiali Zeng, Linfeng Song, Jinsong Su, Jun Xie, Wei Song, and Jiebo Luo. 2020. Neural simile recognition with cyclic multitask learning and local attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9515–9522.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

CCL 2024

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 253-261, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China