Improving the Generalization Ability in Essay Coherence Evaluation Through Monotonic Constraints

Chen Zheng[1,2](✉), Huan Zhang[1], Yan Zhao[1], and Yuxuan Lai[1,2]

[1] The Open University of China, Beijing, China
{zhengchen,zhanghuan,zhaoyan,laiyx}@ouchn.edu.cn
[2] Engineering Research Center of Integration and Application of Digital Learning Technology, Ministry of Education, Beijing, China

Abstract. Coherence is a crucial aspect of evaluating text readability and can be assessed through two primary factors when evaluating an essay in a scoring scenario. The first factor is logical coherence, characterized by the appropriate use of discourse connectives and the establishment of logical relationships between sentences. The second factor is the appropriateness of punctuation, as inappropriate punctuation can lead to confused sentence structure. To address these concerns, we propose a coherence scoring model consisting of a regression model with two feature extractors: a local coherence discriminative model and a punctuation correction model. We employ gradient-boosting regression trees as the regression model and impose monotonicity constraints on the input features. The results show that our proposed model better generalizes unseen data. The model achieved third place in track 1 of NLPCC 2023 shared task 7. Additionally, we briefly introduce our solution for the remaining tracks, which achieves second place for track 2 and first place for both track 3 and track 4.

Keywords: Automated Essay Scoring · Discourse Coherence · Monotonic Constraints

1 Introduction

Discourse coherence refers to the degree to which the various components of a discourse are logically interconnected and contribute to a clear and meaningful message [1]. Analyzing coherence can greatly benefit numerous natural language processing tasks, such as text generation [2], summarization [3] and essay scoring [4,5].

In essay scoring tasks, there are many dimensions to measure the student's language proficiency, such as lexical sophistication, grammatical errors, content coverage and discourse coherence [6]. Since coherence is a key property of a well-written essay, coherence assessment plays an essential role in the task.

F. Liu et al. (Eds.): NLPCC 2023, LNAI 14304, pp. 302–313, 2023.
https://doi.org/10.1007/978-3-031-44699-3_27

In this work, we argue that two key aspects should be considered when evaluating the coherence of an essay. The first aspect is the logical coherence between sentences. The content of the essay should demonstrate a clear progression of ideas, with sentences and paragraphs closely connected and unfolding in logical order. Factors that may negatively impact the logical coherence between sentences include the improper use of discourse connectives and a lack of logical relationships between contexts. The second aspect is the appropriateness of punctuation. Proper punctuation is essential for clarifying the structure and organization of the essay. It can help establish logical connections between sentences, making the text easier to understand. Inappropriate punctuation can lead to confusion and disrupt the smooth flow of the text.

In this work, we propose a feature-based coherence-scoring model framework. We employ two feature extractors to tackle the two essential aspects of coherence. Specifically, the first feature extractor is a local discriminative model [7], while the second is a punctuation correction model [8]. The local discriminative model takes two or three consecutive sentences as input and generates a probability estimate of the local coherence of the sequence. We separated the essay into successive sentences, taking each one as input for the model. Following the inference, we obtained the ratio of coherent sequences to the total number of sequences. The punctuation correction model examines the essay's punctuation usage and explicitly focuses on identifying redundant, missing, and misused commas and periods.

Following feature extractors, we propose employing a regression model to map features onto a final global coherence score. A simple yet transparent model for combining features is linear regression. However, when the patterns in the data exhibit non-linear relationships, alternative models such as random forest regression, gradient-boosted regression trees (GBRT), and neural networks offer superior performance compared to linear regression. A non-linear model may be prone to overfitting the data and negatively impacting the validity of automated scores. To address this issue, we enforce regulations on the input features to maintain linguistically-informed monotonicity, thereby enhancing scoring transparency and improving the model's generalization ability.

Consequently, we present a scoring model that utilizes GBRT and incorporates monotonic constraints on the input features. We assume that the input feature, the ratio of locally coherent sequences to the total sequence of the essay, demonstrates a positive correlation with global coherence. Thus, we apply an increasing constraint to this feature. Furthermore, we assume that the feature of the number of redundant, missing, and misused commas and periods negatively correlates with global coherence. Hence, we impose a decreasing constraint on these features.

In summary, our contributions are as follows:

- We proposed a novel coherence scoring model consisting of a scorer with two feature extractors, i.e. a local discriminative model and a punctuation correction model. We showed that a local discriminative model with a more

extended contextual input performs better than just consecutive pairs of sentences on the subsequent scoring tasks.
- We implement linguistically-informed monotonicity constraints on the input features to enhance the generalization ability in scoring essay coherence.
- Experiments on the LEssay dataset demonstrate the effectiveness of our proposed methods, and we achieved third place on track 1 from NLPCC2023 shared task 7.

In the last of this paper, we will briefly overview our solution for the remaining tracks from NLPCC 2023 shared task 7. The code is available at https://github.com/chernzheng/nlpcc2023_shared_task7_ouchnai_solutions.

2 Related Works

Coherence Modeling. The early development of models for coherence analysis was influenced by lexical cohesion [9], which refers to sharing identical or semantically related words in nearby sentences. Reference [10] introduced the concept of lexical chains and demonstrated that the number and density of lexical chains correlated with the topic structure. Reference [11] introduced the TextTiling algorithm revealing that sentences or paragraphs within a subtopic exhibit higher cosine values than those in neighbouring subtopics. Reference [12]'s LSA Coherence method pioneered the use of embeddings in studying coherence between sentences.

Modern neural representation-learning coherence models [7,13,14] incorporate insights from early unsupervised coherence models for learning sentence representations and assessing their transformations between adjacent sentences. These models are designed to differentiate between natural and unnatural discourses based on deep neural networks.

Automated Chinese Essay Scoring. Reference [15] implemented LDA to score Chinese essays. Reference [16] enhanced the accuracy of Chinese AES by recognizing beautiful sentences and incorporating them as literary features. Reference [17] assessed the organizational score of high school argumentative essays. Reference [18] investigated cross-prompt holistic scoring on four distinct essay sets, with articles in each dataset responding to a distinct prompt. Reference [19] proposed a multi-task learning framework for the Chinese AES and an intersequence attention mechanism to enhance information interaction between the different trait tasks.

3 Method

The architecture of our coherence scoring model is presented in Fig. 1. The model consists of three components: a local discriminative model, a punctuation correction model, and a scorer. The local discriminative model is employed to evaluate the local coherence of consecutive sentences of the essay. The punctuation correction model is utilized to identify the inappropriateness of punctuation usage. The scorer maps the features extracted from the above two models into a final coherence score of the essay.

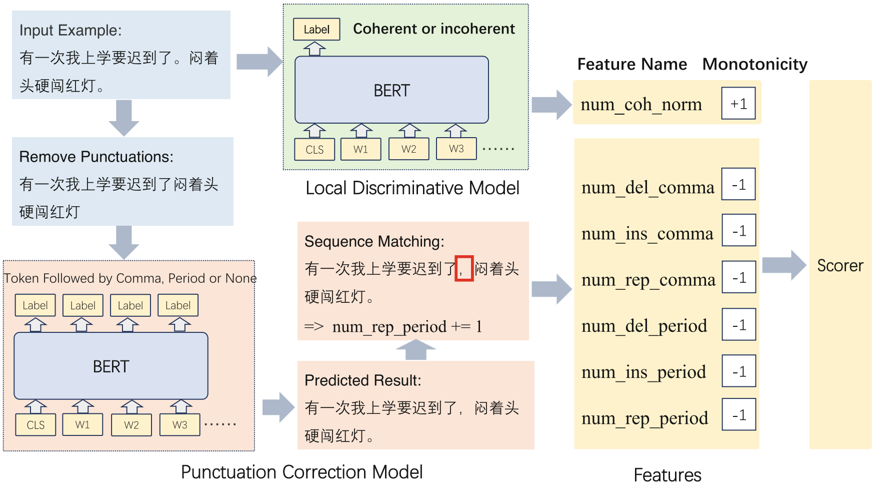

Fig. 1. The figure shows the architecture of our coherence scoring model. The punctuation correction model outputs six features: `num_del_comma`, `num_ins_comma`, `num_rep_comma`, `num_del_period`, `num_ins_period`, and `num_rep_period`, which enforced decreasing constraints on the subsequent scoring process. The local discriminative model output one feature: `num_coh_norm`, which enforces an increasing constraint.

3.1 Local Discriminative Model

Our local discriminative model is similar to that of Ref. [7], but we employ BERT as an encoder and treat the problem as a text classification task. Reference [7] proposed a scoring model to differentiate between consecutive sentence pairs in the training corpus, which are assumed to be coherent, and constructed incoherent ones. We extend the input sequence to three consecutive sentences rather than just two sentences and compare the different context lengths on the performance of subsequent scoring tasks.

For the case of sentence pairs, the input sequence is represented as `[CLS]` + `Sentence A` + `[SEP]` + `Sentence B`, where segment embeddings distinguish between the two sentences. For an essay with n sentences, s_i is the i-th sentence. We construct negative training samples by replacing one of the sentences, s_i or s_{i+1}, with another sentence, s_j $(j \neq i, i+1)$, from the same essay. The trained model denoted as **LD-Bisent**).

For the case of three sentences, the input sequence is set as `[CLS]` + Sentence `A` + Sentence `B` + Sentence `C` without using a special token `[SEP]` to separate them. We randomly substitute one sentence, s_i, s_{i+1} or s_{i+2}, by s_j $(j \neq i, i+1, i+2)$ from the same essay as the negative training sample. The trained model denoted as **LD-Trisent**).

The model use the final hidden vector $C \in R^H$ (in our case, Chinese-RoBERTa-wwm-ext-large [20], H=1024) corresponding to the first input token `[CLS]` as the aggregate representation. The classification layer weights $W \in R^{K \times H}$, where K is the number of labels. In our case, $K = 2$ for coherent or incoherent sequence. We compute a standard classification loss as $\log(\texttt{softmax}(CW^T))$.

3.2 Punctuation Correction Model

Our punctuation correction model is composed of two components. The first component, called the punctuation restoration model, accepts punctuation-free input texts and predicts the label for each token, indicating the punctuation that should follow it. The possible labels include a comma, a period, or no punctuation following the token. The second component is a misused-case classifier, which compares the punctuation-restored text with its original counterpart and determines the type of error the author has made. For instance, consider the sentence written by the author:

有一次我上学要迟到了。闷着头硬闯红灯。

(I ran late for school one day and recklessly charged through the red light.)

To begin with, we remove the punctuation, resulting in the sentence

有一次我上学要迟到了闷着头硬闯红灯

Next, we input this sentence into the punctuation restoration model. The model predicts that the token '了' should be followed by a comma, the token '灯' should be followed by a period, and no punctuation following the rest of the token. Consequently, the punctuation-restored sentence becomes

有一次我上学要迟到了，闷着头硬闯红灯。

Subsequently, the misused-case classifier aligns the punctuation-restored sentence with its original counterpart and identifies that a comma has been erroneously used after the token '了'.

The punctuation restoration model is built upon a token classification model. We remove all punctuation marks from the original text and then pass it through a BERT encoder to obtain the final hidden vector for each input token $T_i \in R^H$. The probability of the token i belonging to one of the labels {0, 1, 2} is computed as $\texttt{softmax}\ (S \cdot T_i)$, where $S \in R^{K \times H}$ is the set of weights to be learned of the final layer. Here, label 0 signifies that the token is not followed by punctuation, label 1 indicates a comma follows it, label 2 indicates it is followed by a period, and $K = 3$ is the number of labels.

The misused-case classifier uses a sequence-matching algorithm to compare the punctuation-restored texts with their original counterparts. We then count the instances of redundant, missing, and misused punctuation in the essay. For the sake of simplicity, all colons within the dataset are transformed into commas.

Semicolons, question marks, and exclamation marks are replaced with periods while disregarding other punctuations.

3.3 Scorer

The scorer takes extracted features from the above two models as input. The one feature is the ratio of coherent sequences to the total number of sequences in the essay (`num_coh_norm`). Additional features are the number of redundant, missing, and misused commas (`num_del_comma`, `num_ins_comma`, and `num_rep_comma`) and the period counterparts (`num_del_period`, `num_ins_period`, and `num_rep_period`).

We employ the abovementioned features as input and utilize a GBRT scorer with monotonic constraints to map these features into a final global coherence score. We impose a decreasing constraint for all features extracted from the punctuation correction model because these features characterize the inappropriateness of punctuation. For feature $x_i \in \{$`num_del_comma`, `num_ins_comma`, `num_rep_comma`, `num_del_period`, `num_ins_period`, `num_rep_period`$\}$, the model satisfies

$$\mathrm{GBRT}(x_1, \ldots, x_i, \ldots, x_n) \geq \mathrm{GBRT}(x_1, \ldots, x_i', \ldots, x_n) \tag{1}$$

whenever $x_i \leq x_i'$. We impose an increasing constraint for feature $x_j =$ `num_coh_norm` because the feature captures the local coherence between adjacent sentences. It satisfies

$$\mathrm{GBRT}(x_1, \ldots, x_j, \ldots, x_n) \leq \mathrm{GBRT}(x_1, \ldots, x_j', \ldots, x_n) \tag{2}$$

whenever $x_j \leq x_j'$.

We compare our proposed scoring model against two regression models: a linear model and a random forest model. We also compare the performance of our model with different configurations, i.e. the scorer with or without monotonic constraints and the local discriminative model with different context lengths.

4 Experiments

4.1 Datasets

LEssay Dataset. The LEssay dataset consists of four sub-datasets corresponding to four tasks. All tasks are related to the coherence evaluation of Chinese student essays. The first sub-dataset is dedicated to the task of global coherence evaluation. It includes a training set of 50 essays, a verification set of 10 essays, and a test set of 5,000 essays. All of these essays are written in Chinese by middle school students and assessed for their coherence on three levels: excellent, moderate, and poor. The remaining three sub-datasets are allocated to the topic sentence extraction, paragraph and sentence logical relation recognition tasks, respectively.

These four tasks are interconnected, and a model trained on one sub-dataset can potentially contribute to another task. However, in this study, a global coherence scoring model will be trained only by the first sub-dataset and two external datasets. These external datasets, including the Chinese essay dataset for pre-training [18] and the IWSLT 2012-zh dataset for punctuation restoration [21], will be utilized to train the feature extractors for the scoring model. The global coherence scores of the first sub-dataset will be used to train the scorer.

Chinese Essay Dataset for Pre-training. The dataset comprises 93,002 essays authored by Chinese students in grades 7 to 12, covering various topics and genres, such as narrative, argumentative, and expository essays.

We utilized the dataset for training the local discriminative model. In practice, we excluded essays with the lowest rating (assigned rating 1) due to poor writing quality. For the remaining essays, we divided each into consecutive sentence pairs or triple sentences, assuming their coherence. And we constructed incoherent sentences, as described in Sect. 3.1. We generated 4.3 million positive and equal negative training samples for the LD-Bisent. We also prepared 3.1 million positive and equal negative training samples for the LD-Trisent.

IWSLT2012-Zh Dataset. The dataset consists of 150k lines of sentences in Chinese from TED talk transcripts. We only predict commas and periods. The question marks are converted to periods for simplicity.

4.2 Experimental Settings

We use the pre-trained Chinese-RoBERTa-wwm-ext-large model to fine-tune the local discriminative and punctuation correction models. For the random forest scorer, we set the number of trees in the forest to 30 and maintained the other parameters at their default values. For the GBRT scorers, we configure the number of boosted gradients to 30, with a maximum tree depth for base learners of 4. The learning rate is set to 1, and all other parameters are left at their default values.

We use precision, recall, and macro F1-score to evaluate the effectiveness of coherence identification. The precision is calculated by dividing the number of correctly identified coherence types (excellent, moderate, and poor) by the total number of identified coherence types. The recall is determined by dividing the number of correctly identified coherence types by the total number of coherence types as labelled.

4.3 Results

Table 1 presents the results of each regression model. In the experiment, we used the LD-Trisent feature extractor in linear and random forest regressions.

Table 1. Comparison of regression models

Model	Precision	Recall	Macro F1
Linear Regression	35.55	48.44	25.57
Random Forest Regression	38.86	23.44	28.74
GBRT (Bi-sent)	33.41	34.10	31.82
GBRT w/ MC (Bi-sent)	36.98	23.02	26.67
GBRT (Tri-sent)	35.77	36.26	34.52
GBRT w/ MC (Tri-sent)	37.28	39.90	33.02

Our findings suggest that the GBRT model with monotonic constraint using LD-Trisent (GBRT w/ MC (Tri-sent)) performs better in terms of precision and recall compared to the same model without enforcing monotonic constraint (GBRT (Tri-sent)). Furthermore, this model demonstrates improvements in precision, recall, and macro F1 score compared to the same model using LD-Bisent (GBRT w/ MC (Bi-sent)) and LD-Bisent without enforcing monotonic constraint (GBRT (Bi-sent)). Additionally, this model exhibits superior performance in macro F1 score compared to both linear and random forest regressions.

Our results show that training local coherence models to predict longer contexts than just consecutive pairs of sentences can result in better performance on subsequent scoring tasks, which agrees with the previous study on discourse representation [22].

5 Our Solution to the Remaining Tracks from NLPCC2023 Shared Task7

5.1 Text Topic Extraction (Track 2)

This task aims to identify the topic sentence for each paragraph and one overall topic sentence for a given middle school student essay.

In our approach, we employ two token classification models to identify both paragraph-level and overall topic sentences. The first model accepts the essay title connected to a paragraph as input. For each token, it outputs a label indicating whether the token belongs to the topic sentences of the paragraph (designated as a key token). The topic sentences of each paragraph are determined by the ratio of key tokens to the total number of tokens within the sentence. We select the sentence with the highest ratio as the topic sentence for that paragraph. The model is fine-tuned on Chinese-RoBERTa-wwm-ext-large.

The second model is similar to the first, but the input is a sequence that sequentially connects the essay title to all paragraph's topic sentences. We assume that the overall topic sentence is one of the paragraph topic sentences and determine it by calculating the ratio of key tokens to the total number of tokens within each paragraph topic sentence. We select the sentence with the highest ratio as the overall topic sentence. The second model is fine-tuned on the first model.

The evaluation results are shown in Table 2. Our approach achieved second place in Track 2.

Table 2. The result of text topic extraction.

Team	Para. Acc.	Full Acc.	Final Acc.	Para. Simi.	Full Simi.
wuwuwu	61.27	34.92	42.82	87.34	80.37
Ours	62.61	33.33	42.12	85.20	79.16

5.2 Paragraph Logical Relation Recognition (Track 3)

The task aims to determine the logical relationship between the two consecutive paragraphs of an essay. The logical relationship includes co-occurrence, inversion, explanatory and superior-subordinate relationships.

Our approach regards the paragraph-level logical relation recognition task as a sequence classification problem. Specifically, we process a pair of paragraphs as input, and the model determines the logical relationship between these paragraphs. Considering the similarity between this task and sentence-level logical relation recognition, we chose to fine-tune the model trained for track 4.

The evaluation results for track 3 are shown in Table 3. Our approach achieved first place in the track.

Table 3. The results of paragraph-level logical relation recognition.

Team	Precision	Recall	Macro F1
Ours	54.66	52.45	52.16
wuwuwu	29.26	28.98	28.77
Lrt123	28.19	30.26	27.54
BLCU_teamworkers	27.17	27.65	25.95

5.3 Sentence Logical Relation Recognition (Track 4)

The task is comparable to the previous task. Nonetheless, the logical relationships are sentence-based and include 12 different relationships.

We employ a two-stage training approach for our classification model. In the first stage, we utilize an external dataset, TED-CDB [23], to pre-train the model based on Chinese-RoBERTa-wwm-ext-large. In the subsequent stage, we fine-tune the pre-trained model on the current dataset to enhance its performance for the given task.

The evaluation results for track 4 are shown in Table 4. Our approach achieved first place in the track.

Table 4. The results of sentence-level logical relation recognition.

Team	Precision	Recall	Macro F1
Ours	36.63	36.36	34.38
wuwuwu	23.49	25.37	23.67
BLCU_teamworkers	7.55	6.30	6.32

6 Conclusion and Future Work

In this study, we present a scoring model to assess the global coherence of Chinese student essays. This scoring model incorporates two feature extractors: a local coherence discriminative model and a punctuation correction model. Furthermore, we employed a GBRT model with linguistically-informed monotonicity constraints to convert features into a final global coherence score.

Our findings suggest that the enforced regulations on the features improved the model's generalization capability, and a local discriminative model with a context extending beyond consecutive sentence pairs can achieve better performance in scoring tasks.

For future research, we will incorporate the features of paragraph-level coherence into the scoring model. The current model considers sentence-level coherence by introducing a local discriminative model. But the global coherence characterized by logical relationships between paragraphs is equally important for coherence evaluation. By incorporating paragraph-level coherence features, we can further enhance the performance of the scoring model and provide a more accurate assessment.

Acknowledgements. This work is supported by NSFC (62206070), the Innovation Fund Project of the Engineering Research Center of Integration and Application of Digital Learning Technology, Ministry of Education (1221014, 1221052), and National Key R&D Program of China (2021YFF0901005).

References

1. Jurafsky, D., Martin, J. H.: Speech and Language Processing, 3rd edn. (Draft of Jan 7, 2023) (2023)

2. Huang, L., Ye, Z., Qin, J., Lin, L., Liang, X.: GRADE: automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 9230–9240 (2020)
3. Christensen, J., Soderland, S., Etzioni, O.: Towards coherent multi-document summarization. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1163–1173 (2013)
4. Miltsakaki, E., Kukich, K.: Evaluation of text coherence for electronic essay scoring systems. Nat. Lang. Eng. **10**(1), 25–55 (2004)
5. Burstein, J., Tetreault, J., Andreyev, S.: Using entity-based features to model coherence in student essays. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 681–684 (2010)
6. Cahill, A., Evanini, K.: Natural language processing for writing and speaking. In: Handbook of Automated Scoring, pp. 69–92. Chapman and Hall/CRC, Boca Raton (2020)
7. Xu, P., et al.: A cross-domain transferable neural coherence model. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 678–687 (2019)
8. Zhang, H., et al.: PaddleSpeech: an easy-to-use all-in-one speech toolkit. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: System Demonstrations, pp. 114–123 (2022)
9. Halliday, M. A. K., Hasan, R.: Cohesion in English (No. 9). Routledge (1976)
10. Morris, J., Hirst, G.: Lexical cohesion computed by thesaural relations as an indicator of the structure of text. Comput. Linguist. **17**(1), 21–48 (1991)
11. Hearst, M.A.: Text Tiling: segmenting text into multi-paragraph subtopic passages. Comput. Linguist. **23**(1), 33–64 (1997)
12. Foltz, P.W., Kintsch, W., Landauer, T.K.: The measurement of textual coherence with latent semantic analysis. Discourse Process. **25**(2–3), 285–307 (1998)
13. Li, J., Li, R., Hovy, E.: Recursive deep models for discourse parsing. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 2061–2069 (2014)
14. Mesgar, M., Strube, M.: A neural local coherence model for text quality assessment. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 4328–4339 (2018)
15. Zhang, M., Hao, S., Xu, Y., Ke, D., Peng, H.: Automated essay scoring using incremental latent semantic analysis. J. Softw. **9**(2), 429–436 (2014)
16. Fu, R., Wang, D., Wang, S., Hu, G., Liu, T.: Elegart sentence recognition for automated essay scoring. J. Chin. Inf. Process. **32**(6), 10 (2018)
17. Song, W., Song, Z., Liu, L., Fu, R.: Hierarchical multi-task learning for organization evaluation of argumentative student essays. In: IJCAI, pp. 3875–3881 (2020)
18. Song, W., Zhang, K., Fu, R., Liu, L., Liu, T., Cheng, M.: Multi-stage pre-training for automated Chinese essay scoring. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 6723–6733 (2020)
19. He, Y., Jiang, F., Chu, X., Li, P.: Automated Chinese essay scoring from multiple traits. In: Proceedings of the 29th International Conference on Computational Linguistics, pp. 3007–3016 (2022)

20. Cui, Y., Che, W., Liu, T., Qin, B., Wang, S., Hu, G.: Revisiting pre-trained models for Chinese natural language processing. Findings Assoc. Comput. Linguist.: EMNLP **2020**, 657–668 (2020)
21. Federico, M., Cettolo, M., Bentivogli, L., Michael, P., Sebastian, S.: Overview of the IWSLT 2012 evaluation campaign. In: Proceedings of the International Workshop on Spoken Language Translation (IWSLT), pp. 12–33 (2012)
22. Iter, D., Guu, K., Lansing, L., Jurafsky, D.: Pretraining with contrastive sentence objectives improves discourse performance of language models. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4859–4870 (2020)
23. Long, W., Webber, B., Xiong, D.: TED-CDB: a large-scale Chinese discourse relation dataset on ted talks. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 2793–2803 (2020)