IFlyEA: A Chinese Essay Assessment System with Automated Rating, Review Generation, and Recommendation

Jiefu Gong[†], Xiao Hu[†], Wei Song[‡], Ruiji Fu[†], Zhichao Sheng[¶],

Bo Zhu[¶], Shijin Wang[†¶£], Ting Liu[§]

[†]State Key Laboratory of Cognitive Intelligence, iFLYTEK Research, Beijing, China
[‡]Academy for Multidisciplinary Studies, Capital Normal University, Beijing, China
[¶]iFLYTEK AI Research (Hefei), China [£]iFLYTEK AI Research (Hebei), LangFang, China
[§]Research Center for SCIR, Harbin Institute of Technology, Harbin, China
{jfgong, xiaohu2, rjfu, zcsheng, bozhu, sjwang3}@iflytek.com,
wsong@cnu.edu.cn, tliu@ir.hit.edu.cn

Abstract

Automated Essay Assessment (AEA) aims to judge students' writing proficiency in an automatic way. This paper presents a Chinese AEA system IFly**E**ssay**A**ssess (IFlyEA), targeting on evaluating essays written by native Chinese students from primary and junior schools. IFlyEA provides multi-level and multi-dimension analytical modules for essay assessment. It has state-of-the-art grammar level analysis techniques, and also integrates components for rhetoric and discourse level analysis, which are important for evaluating native speakers' writing ability, but still challenging and less studied in previous work. Based on the comprehensive analysis, IFlyEA provides application services for essay scoring, review generation, recommendation, and explainable analytical visualization. These services can benefit both teachers and students during the process of writing teaching and learning.

1 Introduction

Automated essay assessment (AEA) is an important educational application (Page, 1968; Rudner et al., 2006). It aims to reduce the burden of teachers for scoring student essays and give students direct instructions to improve their writing ability.

Automated essay scoring (AES) is one of the most important modules for AEA, which is usually formulated as a supervised learning problem. The early approaches utilized hand-crafted features to predict essay scores (Yannakoudakis et al., 2011; Chen and He, 2013; Phandi et al., 2015). Recently, deep learning has been applied to AES as well (Taghipour and Ng, 2016; Dong et al., 2017; Song et al., 2020c).

One issue about AES is that its prediction lacks explainability since a single score gives very limited information. Many efforts have been paid to expand the boundary of AES, and try to analyze detailed linguistic properties, such as grammatical errors (Ng et al., 2014), coherence (Somasundaran et al., 2014), organization (Burstein et al., 2003; Persing et al., 2010) and so on.

Several AES systems, such as E-Rater (Attali and Burstein, 2006) and Lingglewrite (Tsai et al., 2020), have been successfully applied in the education scenario. However, many of them focus on evaluating second-language learners' writing ability or evaluating basic language usages depending on shallow features, which may be not sufficient for evaluating essays written by native speakers. Moreover, most existing platforms mainly target on English, while there are significantly fewer systems working on other languages, such as Chinese.

In this paper, we introduce the IFlyEssayAssess (IFlyEA) system, which is a Chinese automated essay assessment system, focusing on assessing the quality of essays written by native Chinese students from primary and junior schools.

IFlyEA has the following highlights:

- IFlyEA has comprehensive multi-level and multi-dimension analytical modules. It provides state-of-the-art Chinese spelling error correction and grammatical error diagnosis at grammar level. More specially, it also provides rich rhetoric and discourse level analysis, which are less studied but important for evaluating native speakers' writing ability.

- Based on the information provided by the analytical modules, IFlyEA provides a complete set of application services, including rating, review generation and recommendation.

- IFlyEA has an easy-to-use visualization and interactive interface, which can clearly show the detailed analytical results of an essay, and

Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations, pages 240–248, August 1st - August 6th, 2021.

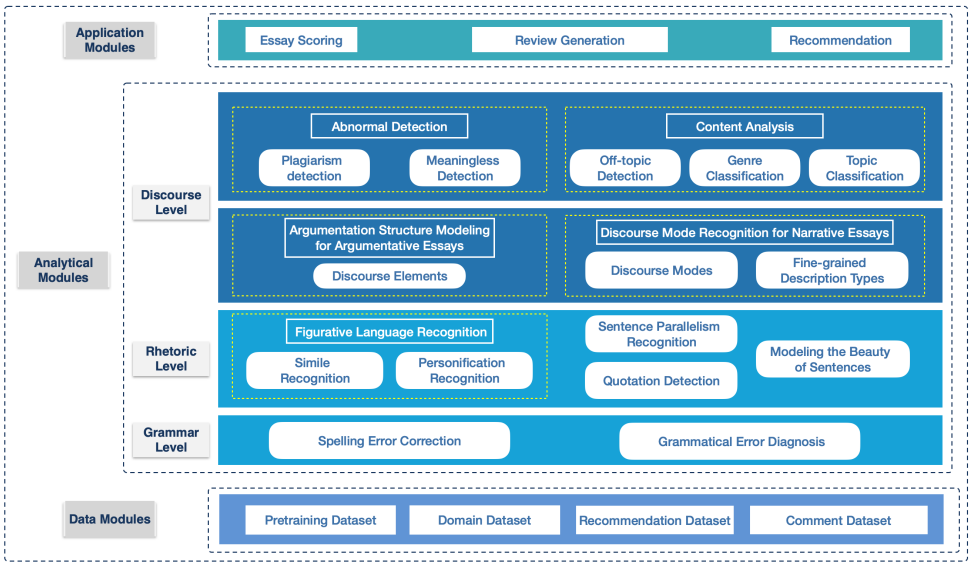

Figure 1: The architecture of IFlyEA.

improve the explainability of predictions at the application level.

The target users of IFlyEA is students from primary and junior schools, in other hands, it is also helpful for teachers to reduce their heavy work. IFlyEA has been applied in practice and it is being continually improved by learning from user feedback.

2 System Architecture

The main modules of IFlyEA can be categorized into two types: **analytical modules** and **application modules**, as shown in Figure 1. These modules are integrated with visualization and interactive interfaces.

The analytical modules involve multi-level and multi-dimension analysis of essay quality, which mainly cover three levels:

- **Grammar level**: This level aims to judge whether students can *correctly* use words to communicate. IFlyEA applies several technical approaches such as spelling correction and grammatical error diagnosis.

- **Rhetoric level**: This level aims to judge whether students can *gracefully* and *skillfully* convey their ideas. IFlyEA can recognize rhetorical devices and *beautiful* sentences in essays.

- **Discourse level**: This level aims to judge whether students can *logically* connect basic discourse units to construct a coherent whole. The system identifies discourse elements for representing and evaluating essay organization, and also has other discourse level analysis such as topic classification and genre classification.

The techniques at grammar level are widely used for essay scoring, especially for evaluating second-language learners. The rhetoric and discourse levels are more important for evaluating essays written by native speakers, especially for distinguishing well-written essays from moderate ones.

The application modules include:

- **Essay scoring**: This module gives scores to indicate the general quality of an essay and the quality of specific aspects.

- **Review generation**: This module provides readable reviews on multiple writing dimensions.

- **Recommendation**: This module suggests relevant and potentially helpful materials to students.

The review generation and recommendation modules depend on the results from the analytical modules and the essay scoring module.

In general, the analytical modules are the basis of the application modules, providing evidence and diagnosis, and also improving the explainability for the predictions of application modules. As illustrated in Figure 2, through web page visualization

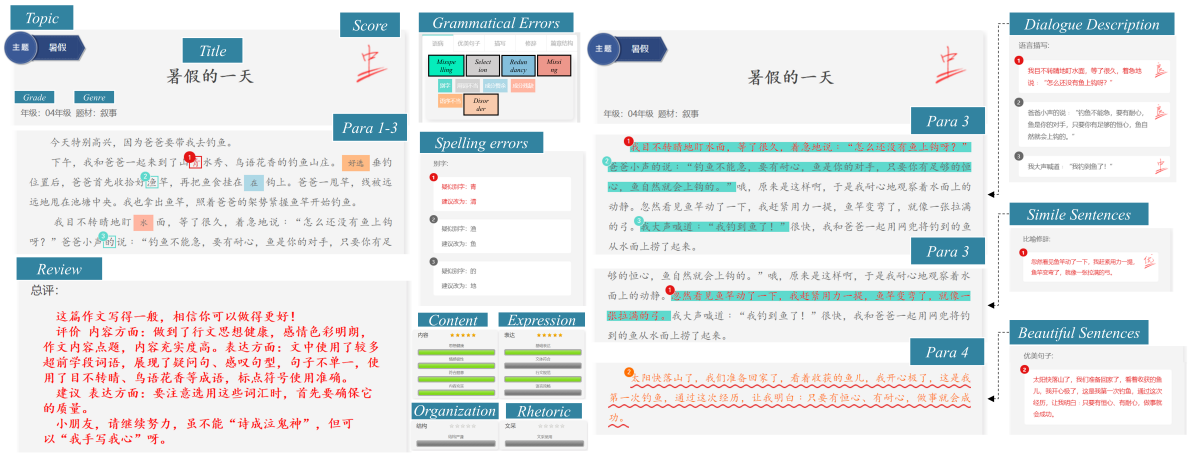

Figure 2: The visualization and interactive interfaces of IFlyEA.

and interfaces, students or teachers can receive rich information and interact with the analytical results.

3 Analytical Modules

IFlyEA has multi-level and multi-dimension quality evaluation to provide comprehensive analytical results. This section will introduce the main analytical modules, which can be roughly categorized into 3 levels: grammar, rhetoric and discourse levels.

3.1 Grammar-level Analysis

Correctly using words is a fundamental requirement for effective writing. Grammar-level analysis would try to detect spelling and grammatical errors in essays, and highlight detected errors as a reminder.

3.1.1 Spelling Error Correction

Given a sentence, our spelling checker would locate spelling errors if there is any, and provide a list of corrected candidates (Tseng et al., 2015).

Inspired by Liu et al. (2013); Yu and Li (2014), we establish a confusion-set based unsupervised two-stage method to detect and correct spelling errors.

Confusion set: A confusion set is built to group characters with similar pronunciation or graphemic into clusters. We implement it with an inverted indexing structure so that given a target character, we can quickly get a list of *confusion characters* from the same cluster.

Stage 1: Correction candidate detection with local context: We train a 5-gram language model `LM` on a large-scale corpus. For each character in a sentence, we substitute it with its corresponding confusion characters one by one, and use `LM` to compute perplexity. If any confusion character leads to a lower perplexity than the original one by a pre-defined threshold, it would be retained as a *correction candidate*. After state 1, we obtain a small list of correction candidates. This stage can be processed very fast.

Stage 2: Correction candidate reranking with global context: We further use the masked language model `MLM` from BERT (Devlin et al., 2019) to take advantage of the pre-trained transformer based language model and exploit the whole sentence as context to rerank the correction candidates at different positions, respectively.

We evaluate our system on the SIGHAN 2015 benchmark. As shown in Table 1, the results demonstrate that our system can obtain competitive results to state-of-the-art methods, although it is unsupervised.

Model	P	R	F_1
Wang et al. (2019)	0.715	0.595	0.649
Zhang et al. (2020)	0.667	0.662	0.664
Ours	0.662	0.641	0.651

Table 1: Chinese spelling error correction performance on SIGHAN 2015 dataset.

3.1.2 Grammatical Error Diagnosis

We focus on 4 types of grammatical errors: *redundant word*, *missing word*, *word selection*, and *word ordering* (Rao et al., 2018). We concentrate on detecting whether a sentence has any grammatical error (detection level), and show the positions of possible grammatical errors (position level).

In line with (Bell et al., 2019; Fu et al., 2018),

Dataset	Detection level	Position level
CGED 2020 data (Rao et al., 2020)	0.894	0.404
Domain data	0.797	0.631

Table 2: Comparison of best F_1 results reported in the CGED 2020 dataset and the domain dataset of primary students' essays.

we formulate grammatical error diagnosis as a sequence labeling problem. Specifically, we build our model based on (Wang et al., 2020), where a ResNet enhanced multi-layer bidirectional transformer encoder (ResELECTRA) is used to encode sentences. This solution ranked 1st in the NLPTEA-2020 CGED shared task at identification and position level.

Since we target on student essays, we continue to train ResELECTRA on a sample of primary students' essays annotated with grammatical error types. The performance on primary students' essays can reach 63% F1-score at position level. The score is higher at position level but lower at detection level than on the CGED 2020 test set. This is because that the label distributions of both levels are different.

3.2 Rhetoric-level Analysis

Grammar-level analysis is important but is not enough for sufficiently evaluating the quality of native speakers' writing. For example, grammatical errors already become much less in junior students' essays compared with that in second-language learners' essays.

This section will introduce rhetoric-level analytical modules, which aim to identify excellent sentences and rhetorical devices, to explore whether language is used in a graceful way.

3.2.1 Modeling the Beauty of Sentences

We define *beautiful sentences* as the ones that can induce aesthetic feelings in us. This definition is vague and the criterion is subjective. Therefore, we construct a classifier to identify beautiful sentences in a data-driven way.

We collect more than $20k$ sentences with beautiful or not labels through crowd-sourcing. Each sentence is at least labeled by two annotators. For training, we only keep the sentences that are labeled with the same tags by two annotators. We train a simple attention based BiLSTM model (Bahdanau et al., 2014) to classify whether a sentence should be annotated as beautiful. The classifier can get an accuracy of 81% through cross-validation evaluation.

3.2.2 Figurative Language Recognition

Figurative language refers to the use of words in a way that deviates from the literal meaning to convey a complicated meaning to amplify our writing. Figurative language recognition in essays enables monitoring students' ability in using figurative language and providing clues for evaluating quality of essays. Currently, we focus on identifying simile and personification.

Simile Recognition Simile leads a comparison between concepts using explicit comparators such as *like, as* in English and *Xiang, Si, Ru* in Chinese. But a sentence with a comparator does not always trigger a simile, unless the two arguments of the comparator form a cross-domain mapping (Lakoff and Johnson, 2008). So simile recognition is not a trivial task.

We adopt a multi-task learning framework for simile recognition (Liu et al., 2018). The framework jointly optimizes two subtasks: *simile sentence classification* and *simile component extraction*. The model is trained on $12k$ annotated sentences that contain a comparator. The simile sentence classifier can obtain a 86% F_1 score in 5-fold cross-validation evaluation on the dataset.

Personification Recognition Personification is another special case of figurative language, borrowing human's actions, expressions, or other characteristics to ascribes specific attributes of non-human objects, such as, "*Life has cheated me*" (Lakoff and Johnson, 2008).

This task is cast as a typical classification problem. We adopt an attention based BiLSTM (Bahdanau et al., 2014) to encode a sentence into a dense feature vector. This vector is then fed into a nonlinear layer and a softmax layer to generate the classification result. Considering the characteristics of this task, we introduce an external knowledge base Chinese CiLin (A Synonymy Thesaurus of Chinese Words) (Mei, 1984) to group words into clusters according to word senses, and assign a learnable embedding vector for each cluster. Each word is represented by the concatenation of its word embedding and cluster embedding, which is fed into the encoder for learning. The personification recognizer can achieve a 80% F_1 score. This task shows to be more difficult than simile recognition.

3.2.3 Sentence Parallelism Recognition

Sentence parallelism is also a widely used rhetorical device in writing. It can be defined as two or more coherent text spans (phrases or sentences), which have similar syntactic structures and related semantics, and express relevant content or emotion together (Song et al., 2016). Parallelism adds balance and rhythm to make speeches and writings more vivid and powerful.

We adopt a feature-based method for this task. The features contain a set of alignment measures at position, word, syntactic and semantic levels. We find that sentence parallelism can be recognized with accepted performance (82% F1-score at pairwise level and 72% F1-score at parallelism block level) using a random forest classifier trained on hundreds of training samples. We also observe that sentence parallelism has a positive correlation to the quality of essays, especially in argumentative essays.

3.2.4 Quotation Detection

Quotation is a figure-of-speech that intentionally referring to some predecessor's words, like poems, maxims, and proverbs, to explain one's own idea, which is aim to amplify the writing or enhance the persuasiveness of argument. We collect a large-scale quotation corpus from the Internet, ranging from poetry to proverbs, and exploit information retrieval (IR) techniques and semantic matching for quotation detection.

3.3 Discourse-level Analysis

Discourse analysis aims to build connections between discourse units to form a whole (Song and Liu, 2020). For essay scoring, we mainly focus on analyzing the organization of essays. One important issue is how to represent essay organization. Our solution is to use discourse elements, which are defined as the function of discourse units in building a coherent discourse. The discourse elements of an essay are dependent on its genre. For example, narrative and argumentative essays usually have different organizational strategies and have different discourse elements.

3.3.1 Argumentation Structure Modeling for Argumentative Essays

For argumentative essays, we define a set of discourse elements following previous work (Attali and Burstein, 2006; Persing et al., 2010), including *prompt*, *thesis*, *main idea*, *support* and *conclusion*. These discourse elements can be used for both sentences and paragraphs (Song et al., 2020a,b).

IFlyEA currently maintains a hybrid organization module. A discourse element is represented by combining its distributed semantic vector and a manually constructed feature vector (Song et al., 2015). The learning framework is based on hierarchical multi-task learning (Song et al., 2020b), which jointly optimizes sentence and paragraph level discourse element identification and organization evaluation. Evaluation shows that some minority discourse elements, such as *thesis* and *ideas*, are more difficult to recognize, and organization evaluation of argumentative essays is still challenging due to the lack of large-scale training data. However, visualizing recognized discourse elements helps teachers quickly see the organization structure of an essay, and helps us collect user feedback through interactions to accumulate more training data.

3.3.2 Discourse Mode Recognition for Narrative Essays

Evaluating organization of narrative essays is even more difficult, since narrative text understanding is still very challenging and open in both theory and practice.

IFlyEA uses discourse modes as discourse elements influenced by (Smith, 2003). The main reasons are: (1) discourse modes can represent the essay organization by segmenting an essay into discourse mode zones; (2) discourse modes are closely related to rhetoric (Connors, 1981; Brooks and Warren, 1958) so that discourse modes can reflect writing proficiency in a degree.

Discourse modes are categorized into *narration*, *description*, *exposition*, *argument* and *emotion*, following (Song et al., 2017). Moreover, we further identify fine-grained description types, such as *appearance, facial expression, action, natural scene, psychology, dialogue* and so on. How to accurately and vividly describe details of a character, a scene or an object is an important lesson to be learned for writing. Identifying and visualizing fine-grained description types let people quickly find some highlights in writing descriptions.

Technically, we adopt a two-stage approach. In the first stage, we use a discourse-level hierarchical encoder to encode an essay and identify 5 discourse modes (Song et al., 2017). The hidden state of each sentence is used as a sentence representation for classification. In the second stage, we further

classify descriptive sentences into fine-grained description types, which is formulated as a typical classification problem.

3.3.3 Discourse-level Abnormal Detection and Content Analysis

Abnormal detection is important for building a robust system. For example, intentional plagiarism is a terrible behavior and should be detected. We build a large-scale corpus covering common plagiarism resource, and exploit IR techniques and semantic matching to detect plagiarism. We also filter out malicious input, such as non-Chinese essays or meaningless character sequences, utilizing a pre-trained language model.

Other content analysis, including off-topic detection, genre classification, and topic classification, are also required to support the comprehensive assessment of essays. We formulate these tasks as a classification problem. The genre and topic classification can be well solved, while off-topic detection is very challenging at present.

4 Application Modules

4.1 Essay Scoring

Essay scoring is a main module for AES. Instead of giving a single general score only, we consider scoring from multiple aspects additionally, including *content*, *expression*, *rhetoric* and *organization*, to provide a comprehensive assessment.

We formulate these scoring tasks as an essay classification problem, classifying a given essay into four grades: *bad, moderate, good* and *excellent*. We construct a feature-based model for each task, and use different feature templates for different aspects. The feature templates can be divided into three types: *basic features*, such as length, vocabulary, syntax and distributed dense representations; *common analytical features*, which are based on the output of our analytical modules, such as the counts of spelling and grammatical errors, and the use of rhetorical devices; and *genre related features*, for example, we use different strategies for modeling the organization of narrative and argumentative essays so that the features would be extracted accordingly.

4.2 Review Generation

Generating a review based on the multi-level evaluation can benefit students for getting direct instructions, and also benefit teachers for getting scoring reports fast and automatically. Currently, our system generates reviews based on a series of predefined templates. The scores of multiple aspects and the whole essay are generated by essay scoring module. According to these scores, the system would manage template selection and integration to generate a coherent review, revealing both the advantages and the shortcomings of an essay.

4.3 Recommendation

In addition to rate and review essays, it is also important to help students learn from feedback to overcome existing weaknesses. To tackle this, We build a module to recommend relevant materials according to diagnosis results at three levels.

We trigger the grammar-level recommendation if spelling errors are detected. In addition to recommending the correct characters, IFlyEA will automatically generate a set of cloze test questions. We first retrieve sentences containing the correct character from an existing corpus of this module, then mask the character in each sentence, and mix it with characters from its confusion set, and finally let students choose the best character to fill the blank. We expect students can better master correct usage of characters and distinguish confusion characters through exercises. As a supplement, the meaning and example usage of both the correct character and its confusion set are prepared previously, which will be displayed after the exercises.

At rhetoric level, we recommend some well-written rhetorical sentences that describe similar objects or scenes as in the target essay, while at discourse level, we show more well-written essays or passages related to similar topics. To achieve this, we have constructed a high quality resource bank of high scoring essays, proses and novels written by famous writers. We use the analytical modules to analyze the resource to support recommendations according to different demands.

5 Conclusion and Future Work

This paper presented IFlyEA, a Chinese automated essay assessment system. IFlyEA demonstrates the techniques, that we have developed, could tackle with evaluating the quality of essays written by native Chinese students. A demonstrating video is available at **https://youtu.be/BujBQfxvX3A**.

The main advantage of IFlyEA is its multi-level and multi-dimension analytical modules for essay assessment, especially on several high level skill-

ful language usage abilities, which is less studied previously. Most of these modules can achieve moderate and above performance. IFlyEA also provides comprehensive services for rating, review generation and recommendation. Together with the visualization and interactive interfaces, teachers and students can get useful feedback and easily understand why the system makes such predictions.

IFlyEA has been applied in practice. In future, we plan to conduct more user studies and continue to improve the system. And how to evaluate the impact of the system on students is another important problem, which is worth exploring.

References

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater v. 2. *The Journal of Technology, Learning and Assessment*, 4(3).

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Samuel Bell, Helen Yannakoudakis, and Marek Rei. 2019. Context is key: Grammatical error detection with contextual word representations. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 103–115, Florence, Italy. Association for Computational Linguistics.

Cleanth Brooks and Robert Penn Warren. 1958. *Modern rhetoric*. Harcourt, Brace.

Jill Burstein, Daniel Marcu, and Kevin Knight. 2003. Finding the write stuff: Automatic identification of discourse structure in student essays. *IEEE Intelligent Systems*, 18(1):32–39.

Hongbo Chen and Ben He. 2013. Automated essay scoring by maximizing human-machine agreement. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1741–1752, Seattle, Washington, USA. Association for Computational Linguistics.

Robert J Connors. 1981. The rise and fall of the modes of discourse. *College Composition and Communication*, 32(4):444–455.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Fei Dong, Yue Zhang, and Jie Yang. 2017. Attention-based recurrent convolutional neural network for automatic essay scoring. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 153–162.

Ruiji Fu, Zhengqi Pei, Jiefu Gong, Wei Song, Dechuan Teng, Wanxiang Che, Shijin Wang, Guoping Hu, and Ting Liu. 2018. Chinese grammatical error diagnosis using statistical and prior knowledge driven features with probabilistic ensemble enhancement. In *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 52–59, Melbourne, Australia. Association for Computational Linguistics.

George Lakoff and Mark Johnson. 2008. *Metaphors we live by*. University of Chicago press.

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, Brussels, Belgium. Association for Computational Linguistics.

Xiaodong Liu, Kevin Cheng, Yanyan Luo, Kevin Duh, and Yuji Matsumoto. 2013. A hybrid Chinese spelling correction using language model and statistical machine translation with reranking. In *Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing*, pages 54–58, Nagoya, Japan. Asian Federation of Natural Language Processing.

Jiaju Mei. 1984. 同义词词林*(Synonymy Thesaurus of Chinese Words)*, volume 1983. 商务印书馆; 上海.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The CoNLL-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14, Baltimore, Maryland. Association for Computational Linguistics.

Ellis B Page. 1968. The use of the computer in analyzing student essays. *International review of education*, pages 210–225.

Isaac Persing, Alan Davis, and Vincent Ng. 2010. Modeling organization in student essays. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 229–239.

Peter Phandi, Kian Ming A. Chai, and Hwee Tou Ng. 2015. Flexible domain adaptation for automated essay scoring using correlated linear regression. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 431–439, Lisbon, Portugal. Association for Computational Linguistics.

Gaoqi Rao, Qi Gong, Baolin Zhang, and Endong Xun. 2018. Overview of NLPTEA-2018 share task Chinese grammatical error diagnosis. In *Proceedings*

of the 5th Workshop on Natural Language Processing Techniques for Educational Applications, pages 42–51, Melbourne, Australia. Association for Computational Linguistics.

Gaoqi Rao, Erhong Yang, and Baolin Zhang. 2020. Overview of NLPTEA-2020 shared task for Chinese grammatical error diagnosis. In *Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 25–35, Suzhou, China. Association for Computational Linguistics.

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

Carlota S Smith. 2003. *Modes of discourse: The local structure of texts*, volume 103. Cambridge University Press.

Swapna Somasundaran, Jill Burstein, and Martin Chodorow. 2014. Lexical chaining for measuring discourse coherence quality in test-taker essays. In *Proceedings of COLING 2014, the 25th International conference on computational linguistics: Technical papers*, pages 950–961.

Wei Song, Ruiji Fu, Lizhen Liu, and Ting Liu. 2015. Discourse element identification in student essays based on global and local cohesion. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2255–2261, Lisbon, Portugal. Association for Computational Linguistics.

Wei Song and LiZhen Liu. 2020. Representation learning in discourse parsing: A survey. *Science China Technological Sciences*, pages 1–26.

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, Osaka, Japan. The COLING 2016 Organizing Committee.

Wei Song, Ziyao Song, Ruiji Fu, Lizhen Liu, Miaomiao Cheng, and Ting Liu. 2020a. Discourse self-attention for discourse element identification in argumentative student essays. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2820–2830, Online. Association for Computational Linguistics.

Wei Song, Ziyao Song, Lizhen Liu, and Ruiji Fu. 2020b. Hierarchical multi-task learning for organization evaluation of argumentative student essays. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3875–3881. International Joint Conferences on Artificial Intelligence Organization. Main track.

Wei Song, Dong Wang, Ruiji Fu, Lizhen Liu, Ting Liu, and Guoping Hu. 2017. Discourse mode identification in essays. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 112–122.

Wei Song, Kai Zhang, Ruiji Fu, Lizhen Liu, Ting Liu, and Miaomiao Cheng. 2020c. Multi-stage pre-training for automated chinese essay scoring. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6723–6733.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891, Austin, Texas. Association for Computational Linguistics.

Chung-Ting Tsai, Jhih-Jie Chen, Ching-Yu Yang, and Jason S. Chang. 2020. LinggleWrite: a coaching system for essay writing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 127–133, Online. Association for Computational Linguistics.

Yuen-Hsien Tseng, Lung-Hao Lee, Li-Ping Chang, and Hsin-Hsi Chen. 2015. Introduction to SIGHAN 2015 bake-off for Chinese spelling check. In *Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing*, pages 32–37, Beijing, China. Association for Computational Linguistics.

Dingmin Wang, Yi Tay, and Li Zhong. 2019. Confusionset-guided pointer networks for Chinese spelling check. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5780–5785, Florence, Italy. Association for Computational Linguistics.

Shaolei Wang, Baoxin Wang, Jiefu Gong, Zhongyuan Wang, Xiao Hu, Xingyi Duan, Zizhuo Shen, Gang Yue, Ruiji Fu, Dayong Wu, Wanxiang Che, Shijin Wang, Guoping Hu, and Ting Liu. 2020. Combining ResNet and transformer for Chinese grammatical error diagnosis. In *Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 36–43, Suzhou, China. Association for Computational Linguistics.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189, Portland, Oregon, USA. Association for Computational Linguistics.

Junjie Yu and Zhenghua Li. 2014. Chinese spelling error detection and correction based on language model, pronunciation, and shape. In *Proceedings of The Third CIPS-SIGHAN Joint Conference on Chinese Language Processing*, pages 220–223, Wuhan, China. Association for Computational Linguistics.

Shaohua Zhang, Haoran Huang, Jicong Liu, and Hang Li. 2020. Spelling error correction with soft-masked BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 882–890, Online. Association for Computational Linguistics.