Natural Language Engineering (2022), **28**, pp. 449–486
doi:10.1017/S1351324921000103

CAMBRIDGE
UNIVERSITY PRESS

ARTICLE

Automated evaluation of the quality of ideas in compositions based on concept maps

Li-Ping Yang[1], Tao Xin[1,*], Fang Luo[2], Sheng Zhang[1] and Xue-Tao Tian[3]

[1]Collaborative Innovation Center of Assessment for Basic Education Quality, Beijing Normal University, Beijing, P. R. China, [2]Department of Psychology, Beijing Normal University, Beijing, P. R. China and [3]School of Computer and Information Technology, Beijing Jiaotong University, Beijing, P. R. China
*Corresponding author. E-mails: xintao@bnu.edu.cn

(Received 12 September 2019; revised 26 April 2021; accepted 26 April 2021; first published online 24 May 2021)

Abstract

Nowadays, automated essay evaluation (AEE) systems play an important role in evaluating essays and have been successfully used in large-scale writing assessments. However, existing AEE systems mostly focus on grammar or shallow content measurements rather than higher-order traits such as ideas. This paper proposes a new formulation of graph-based features for concept maps using word embeddings to evaluate the quality of ideas for Chinese compositions. The concept map derived from the student's composition is composed of the concepts appearing in the essay and the co-occurrence relationship between the concepts. By utilizing real compositions written by eighth-grade students from a large-scale assessment, the scoring accuracy of the computer evaluation system (named AECC-I: Automated Evaluation for Chinese Compositions—Ideas) is higher than the baselines. The results indicate that the proposed method deepens the construct-relevant coverage of automatic ideas evaluation in compositions and that it can provide constructive feedback for students.

Keywords: Automated essay scoring; Ideas; Chinese compositions; Concept maps; Writing ability

1. Introduction

Automated essay evaluation (AEE) is the process of evaluating and scoring written essays via computer programs using features intended to measure many of the traits specified in scoring rubrics, such as the six-trait model (Spandel and Stiggins 1990).[a] Nowadays, AEE systems play an important role in writing evaluation and have been successfully used in large-scale writing assessments. Automated assessment holds the potential for maintaining scoring reliability and desired timelines for reporting scores in large-scale assessments (Ramineni and Williamson 2013). A large body of data shows that the reliability of AEE scores is superior to human ratings (Attali and Burstein 2006; Shermis *et al.* 2010; Burstein, Tetreault and Madnani 2013).

1.1 The trait of ideas in an essay

The trait of ideas in an essay is considered one of the most important characteristics in the existing essay scoring rubric (Spandel and Stiggins 1990; Cui 2001; the National Writing Project,

[a]The six-trait scoring model (Spandel et al. 1990) has garnered a following in the educational community (Quinlan et al. 2009) focusing on ideas, organization, voice, word choice, sentence fluency, and conventions. After the seventh trait called presentation was developed, the model changed its name to the 6+1 trait scoring model.

NPC, America 2006; Ruegg and Sugiyama 2013), which reflects the writing construct. The quality of ideas includes the main idea, details, information and material, reasoning, and the selection and confirmation of evidence (Northwest Regional Educational Laboratory, NWREL 2014). According to the curriculum standards for Standard Chinese language and other subjects in compulsory education (2011) issued by the Ministry of Education of China, the ideas in the composition are often the first consideration in the process of written assessment (Feng 1990; Jin 1992; Cui 2001). The criteria for evaluating compositions according to the assessment of teaching quality for basic education (ATQE) include five basic elements for evaluating compositions (Table 1 contains the scoring rubric).[b] Three aspects to describe the quality of the ideas can be derived from the rubric:

- Is there a main idea?
- Does the content support the main idea?
- Are the ideas developed adequately?

For example, high-scoring essays usually present a clear and well-focused main idea along with specific details to support their main point; whereas, low-scoring essays contain irrelevant content and an unclear main idea that might not even display a connection to the topic.

However, it is worth noting that the construct for AEE system measurements differs from human scoring in rating essays no matter how strongly the computer and human scoring correlate (Deane 2013). AEE systems tend to be more detail-oriented since they usually pay more attention to specific sets of quantitative linguistic features. Most AEE systems generate general quality holistic scores for composition, while some of them also score for different potential composition traits (Shermis 2002; Attali 2011, 2013; Zedelius, Mills and Schooler 2018). The trait of content is easier to measure than ideas, but both traits are linked (Quinlan, Higgins and Wolff 2009). Current AEE systems extract text features to capture the content characteristics of an essay by employing content vector analysis (e.g., e-rater, Educational Testing Service) (Burstein, Tetreault and Madnani 2013), latent semantic analysis (LSA) (Landauer, Lochbaum and Dooley 2009), or latent Dirichlet allocation methods to compare similarities between essays and then determine the content-level quality. Also, some current AEE engines use different specific features based on statistics or machine learning methods to detect whether an essay is off-topic (Higgins, Burstein and Attali 2006; Louis and Higgins 2010; Chen et al. 2015; Rei 2017).

Research on automatic scoring of Chinese compositions started relatively later than the automatic scoring of English compositions. Some studies measure shallower linguistic aspects, such as the choice of the words, (Chen 2016) and the recognition of beautiful sentences via computer programs (Liu, Qin and Liu 2016; Fu *et al.* 2018). The LSA method was also introduced to assess the similarity to high-scoring Chinese compositions (Cao and Yang 2007). Bochen Kuo's team (in National Taichung University of Education, Taiwan) developed a Chinese Coh-Metrix online text analysis system (Ni *et al.* 2013, 2014; Zhang *et al.* 2014) and Mo (2018) used the features in this system to predict the writing ability of Chinese second language learners.

1.2 Graph-based method for essay scoring

A concept map can be employed to analyze a student's understanding of a complex situation (Koszalka and Spector 2003; Zouaq and Nkambou 2008; Villalon and Calvo 2011; Zouaq, Gasevic and Hatala 2011). Using language as the basis for constructing a concept map is likely to represent a more accurate picture of the meaning and structure of the targeted internal knowledge (Pirnay-Dummer *et al.* 2010; Kim 2012a) because structural knowledge consists of concepts and

[b] The five elements are ideas, content, structure, expression, and presentation.

Table 1. Ideas' scoring guide for Chinese composition in ATQE

Score	Scoring guide for the ideas in a composition
6	Excellent and novelty response: • sustains a well-focused main idea; • expresses ideas precisely and well and effectively addresses the topic; • the ideas may be interesting and creative.
5	Skillful response: • has a focused main idea; • supports the ideas with pertinent content through much of the response; • exhibits some variety in sentences and uses good word choice, and occasionally uses words inaccurately.
4	Sufficient response (maybe marked by one or more of the following): • takes a clear position; • supports the ideas with some pertinent content and developed the ideas to some extent; • sentences and words may be simple and unvaried, but can largely express the topic, with some errors, but they do not interfere with understanding the ideas
3	Insufficient response (maybe marked by one or more of the following): • attempts to take a position to addresses the topic, but the position or main idea is unclear; • provides uneven support for the ideas and no development, or maybe very brief; • the response may be a short or too repetitive and limited expression of the ideas.
2	Unsatisfactory response (maybe marked by one or more of the following): • has a very unclear main idea; • uses irrelevant information in much of the response to the task; • sentences and word choice may often be inaccurate and interfere with understanding the ideas.
1	off-topic: • takes an unfocused position to address the topic; • provides no relevant content in response to the topic; • maybe very few words, or only a beginning and does not respond to the topic; • errors severely impede understanding the response.
0	No answer (maybe marked by one or more of the following): • is blank; • merely copies the topic or other items in the text; • is illegible, nonverbal, or unmeaningful words.

Note: The table is an English translation of the scoring rules (originally in Chinese) for the ideas in ATQE.

relations (Koszalka and Spector 2003). Each essay is transformed into a concept map, in which nodes correspond to words, n-grams, or other tokens of text, and the co-occurring ones are connected by weighted edges. However, the use of concept maps for educational assessment is just beginning. Concept maps are used as a visual inspection to reveal the assessment context and writing style. In performance assessments, a concept map can be used to investigate the abilities of a writer to solve complex problems (Schlomske and Pirnay-Dummer 2008; Kim 2012a, 2012b; Zhou, Luo and Chen 2018), but these studies do not utilize full automation in the assessment process. Somasundaran *et al.* (2016) investigated whether the development of ideas in writing could

https://doi.org/10.1017/S1351324921000103 Published online by Cambridge University Press

be captured by graph properties derived from the text and if this could be used to improve holistic scoring in persuasive and narrative essays. Also, text analysis based on complex networks proves that composition quality has a strong correlation with complex network features (such as the degree of nodes) (Antiqueira *et al.* 2007; Amancio, Oliveira and Costa 2012), which can be used in essay scoring (Ke, Zeng and Luo 2016; Zupanc and Bosnic 2017). The commonality and individuality of these AEE systems indicate a tight-knit relationship between graph-based characteristics and essay quality.

For the construction of a concept map, our method is based on Somasundaran *et al.* (2016) and Zupanc and Bosnic (2017), but there are some differences. Essay words in our study were merged synonymously to enhance the meanings of the nodes and create refined concepts rather than recognizing nodes by word types (Somasundaran *et al.* 2016). Whereas Zupanc and Bosnic (2017) utilized the sliding window method to transform a composition into a continuous token sequence, where each token is represented as a node and only contacts its front and back nodes. Their method is suitable for evaluating composition coherence but is inapt at concept map forming. Besides, Janda *et al.* (2019) used sentences as nodes, and the semantic similarity between these sentences as the weight of the edges to construct a graph, which differs from our concept maps.

In terms of developing graph-based features, we combined graph theory features, and some features adopted from Somasundaran *et al.* (2016) and Zupanc and Bosnic's (2017) studies. However, Somasundaran *et al.* (2016) did not use distributed semantics to represent concept map nodes, whereas our research employed word2vec embedding to represent essay concepts. Although Zupanc and Bosnic (2017) used high-dimensional vectors to represent nodes and calculate features, the meaning of the nodes and edges is quite different from our concept maps. Zupanc and Bosnic's features are used to access the distribution pattern of sequential overlapping essay parts, and our research evaluates the distribution pattern of concepts in an essay. Moreover, Maharjan and Rus (2019) transformed student answers into concept maps and compared the similarity of learning tuples between ideal concept maps and student-made maps to evaluate their answers. This method is suitable for short-answer questions, which usually have a golden (or reference) answer. However, composition semantics are superior with open-ended questions; therefore, we compared the similarity of a targeted essay with a high-scoring essay set to develop features. Finally, past researchers concentrated on general composition assessment, while this study's main objective is to reveal the potential relationship between concept map characteristics and idea traits.

1.3 Argument mining in automatic essay scoring

There is another line of work that applies argumentation mining (AM) technology to the automatic scoring of persuasive essays in the education field. This research is based on identifying argumentative structures that consist of argumentative components such as claims, premises (Mochales-Palau and Moens 2009; Kwon *et al.* 2007; Eckle-Kohler, Kluge and Gurevych 2015; Stab and Gurevych 2014a, 2014b, 2017), and the relationship between these components (Ong, Litman and Brusilovsky 2014; Stab and Gurevych 2014a, 2014b; Ghosh *et al.* 2016; Persing and Ng 2016; Nguyen and Litman 2016; Ghosh, Klebanov and Song 2020; Persing and Ng 2020), which can help represent complex concepts in essays written about controversial topics. There are two scenarios for the use of argument mining technology in the automatic evaluation of persuasive essays. The first is the use of argumentation features to automatically grade particular essay quality dimensions, such as topic relevance, (Persing and Ng 2014), views and objectives (Farra, Somasundaran and Burstein 2015), and argument strength (Persing and Ng 2015; Wachsmuth, Khatib and Stein 2016; Ke 2019). The other is examining the contribution of the argumentative features to the quality of the holistic score (Persing and Ng 2013, 2015; Ong, Litman and Brusilovsky 2014; Song *et al.* 2014; Farra Somasundaran and Burstein 2015; Ghosh *et al.* 2016). Nguyen and Litman (2018) proposed a method for automatically extracting features in the whole process of cross-prompt argumentation and analyzed the effect of each dimension feature using the controlled variable method. Some automatic essay scoring systems are based on scoring rubrics (Rahimi *et al.* 2017;

https://doi.org/10.1017/S1351324921000103 Published online by Cambridge University Press

Zhang and Litman 2017, 2018; Zhang *et al.* 2019), while some utilize end-to-end neural models (Cocarascu and Toni 2018).

Although recent years have seen a surge of interest in automated essay scoring based on AM technology, there are still some problems that inevitably limit the use of AM in essay scoring. Argumentation is an important aspect of persuasive essay assessment, but it is not equally effective for all essay types. For example, the data collected in this study use essays written by eighth-grade Chinese students that were obtained from a large-scale test, and the corpus of each prompt is a mixture of various essay genres such as descriptive, narrative, and persuasive. Most AM-based research on persuasive writing is aimed at college students' writing, so younger students' writings have received insufficient analysis. Due to this knowledge gap, standard discourse structures (such as thesis, main proposition, support, and conclusion) are not shown, so the argument component is rare or difficult to identify (Ghosh, Klebanov and Song 2020) in younger students' writings. Moreover, different prompts often lead to diverse arguments, and few researchers have made progress in scoring general writings (Mayfield and Black 2020). In fact, if the evaluation objective is the ideas generated by the students in the writing process, it should also consider contextual details in addition to an essay's argumentive part. In this regard, researchers have noticed the influence of contextual knowledge on the identification of argumentative structures (Nguyen and Litman 2016; Saint-Dizier 2017; Opitz and Frank 2019); however, these studies did not directly use contextual information to predict the quality of the arguments.

The concept map method in our research can be seen as a promising alternative in overcoming these problems, although it faces challenges in the representation of complex concepts in argumentation. To evaluate the quality of ideas across different genres and essay forms, and not just confined to the concepts that appear in the argumentation, we use concept maps to capture the overall quality of ideas in the global concept distribution of the context. In text computing, there are two different views on coding and representing knowledge. The first way, which is popular in the deep learning community, is to embed the semantics of text into a high-dimensional vector space through pretraining, such as word2vec (Mikolov *et al.* 2013), Glove (Pennington, Socher and Manning 2014), and bidirectional transformer architecture (BERT) (Devlin *et al.* 2019). Furthermore, the knowledge embedded in the pretrained semantic vector is implicit and unstructured. The other way is from a structural point of view. As mentioned above, concept maps (or semantic networks) can reflect the conceptualization space generated by an author, but structural evidence based on graphs has not been fully explored.

This study mapped the student essays to a concept map in high-dimensional semantic space to show the advantage of using concept maps to combine pretrained distributed semantic information and the structure information of a map to predict the idea scores. The goal of this research is to develop an automatic scoring system (named AECC-I: Automated Evaluation for Chinese Compositions-Ideas) for ideas with generalization and to investigate whether the system can capture changes in the quality of the ideas of each subdimension based on a rubric.

2. Capturing the quality of ideas using concept maps

2.1 Essay as a concept map

In this study, it is assumed that a concept map can effectively assess a student's conceptualization of a problem in a writing task, and a concept map elicited from a student's response by computer is used to represent the underlying semantics embedded in the essay.

2.1.1 Preprocessing

Three text-preprocessing steps were implemented by the computer before extracting concept maps: Chinese word segmentation (CWS), part-of-speech (POS) tagging, and filtering the stop word (FSW). To do that, the Language Technology Platform Cloud (LTP-Cloud) is used in this

https://doi.org/10.1017/S1351324921000103 Published online by Cambridge University Press

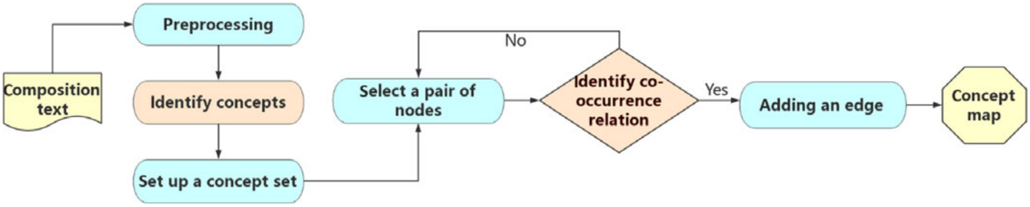

Figure 1. The process of constructing a concept map from a composition.

study,[c] which was developed by the Research Center for Social Computing and Information Retrieval at Harbin Institute of Technology (HIT-SCIR). LTP-Cloud is a widely used Chinese language processing platform that we have completed word segmentation and POS tagging for Chinese compositions before. Steps for text preprocessing are as follows:

Step 1: Chinese word segmentation

Because sentences written in Chinese are composed without spaces and require word segmentation, CWS, a process of segmenting Chinese sentences into word sequences, is required. Since words are the basic units of language that make sense in Chinese, word segmentation is the initial step for information retrieval, text classification, sentiment analysis, and many other Chinese natural language processing tasks.

Step 2: POS tagging

POS is the process of labeling each word in a sentence with its appropriate part of speech (deciding whether each word is a noun, verb, adjective, etc.). Compared with English, Standard Chinese lacks morphological changes, so it cannot be used to identify parts of speech. LTP can solve this problem by using a dictionary query algorithm based on string matching and a POS algorithm based on statistics.

Step 3: Filtering the stop word

To filter noise in word segmentation results, a stop word list made up of 1208 words was created. This filtering removes function words such as “a” as well as punctuations.

2.1.2 Constructing concept maps

Concept maps are automatically constructed from essays by representing each concept in an essay as a node in the concept map. Concepts distilled from an essay are based on a set of rules, and edges are created by connecting all nodes (concepts) to other nodes based on their co-occurrence relation in a sentence. The assumption is that the edges between a pair of concepts simulate the links of a writer’s ideas in an essay. A flowchart of the concept map construction process is shown in Figure 1. Two key steps are described below.

–Identifying concepts

Words are the basic unit of language that carries meaning in Standard Chinese, but some words share the same meaning or very similar meanings. For example, a mobile phone could be contextually synonymous with a cellphone, and they both refer to a common concept in general. In this study, a synonymy merging process is used to collapse similar words based on Cilin.[d] Cilin contains more than 10 million entities, which are divided into 12 categories, 94 middle categories, and

[c] LTP-Cloud URL: http://www.ltp-cloud.com, developed by HIT-SCIR.

[d] Cilin (V2.0). URL: http://www.bigcilin.com was compiled by Mei in 1983 and extended by HIT-SCIR (2018, 2019); it is a Chinese synonym database similar to WordNet.

https://doi.org/10.1017/S1351324921000103 Published online by Cambridge University Press

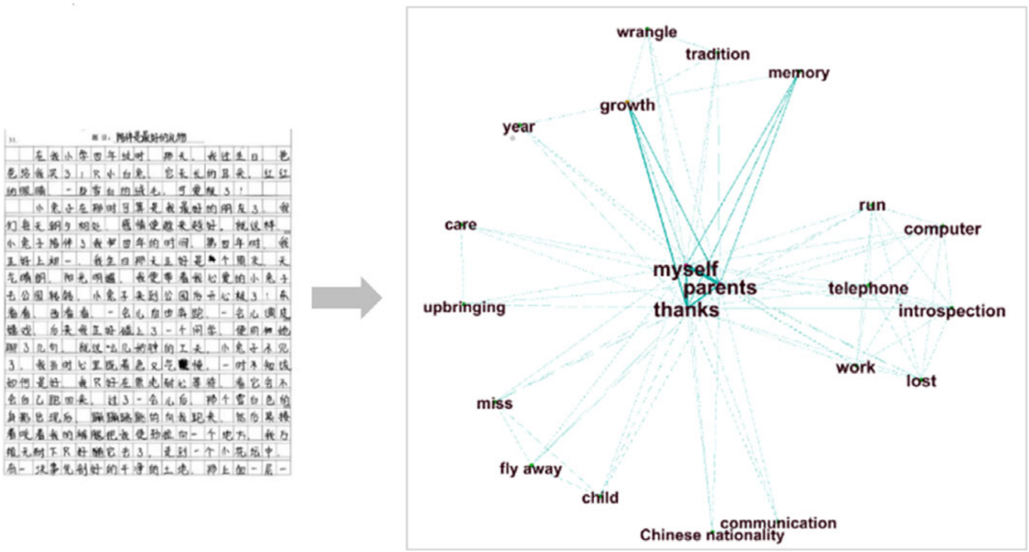

Figure 2. The composition is transformed into a concept map (Chinese characters have been translated into English in the graph on the right.

1428 subcategories. The most detailed level is the atomic word groups; for each synonym atomic word group, a uniform general word is used to represent the concept of the word group. Then, the concepts were distilled from the word groups depending on a set of rules via a computer program. Finally, a concept set obtained through the above process serves as the node set on a concept map, and the number of nodes on a concept map is taken as equal to the number of unique concepts in an essay.

The concepts were distilled from general words depending on a set of modified rules based on Somasundaran *et al.* (2016):

- Rule 1: A concept can be a general noun, compound noun, pre-modifier adjective, direction noun, or a named entity (the name of a person, location, organization, or other entities in a sequence of words recognized by LTP tools).
- Rule 2: Distilled concepts are primarily stored as standardized word types.
- Rule 3: Co-reference resolution, which means pronouns are replaced with the nouns they represent, except for first-person pronouns.

– Identifying co-occurrence relation

Identifying links among concepts is the most important step because it provides information used to construct edges on concept maps. A pair of concepts appearing together usually means there is a syntactic or semantic connection between them (Zhou *et al.* 2008). An example of an essay converted into a concept map is illustrated in Figure 2. The gray lines on the right graph represent the co-occurrence relations between pairs of concepts in this essay.

In order to evaluate the quality of an automatically extracted concept map, two teachers and five students participated in the manual construction of a concept map for each of the 100 randomly selected compositions from the same datasets. Before participating in the evaluation, these teachers and students participated in formal scoring for compositions (described in Section 3.1), so they have adequate familiarity with the writing topics and essay content. While constructing a

Table 2. A framework of features extracted by computer

Types of features	Features	Word embeddings
Global convergence	Moran's I (*MI*)	Y
	Standard distance (*SD*)	Y
	Mean of PageRank value (*MPR*)	
Local convergence	Local Gettis's G (*LG*)	Y
	Number of maximal cliques (*NMC*)	
	Graph transitivity (*GT*)	
Distance between nodes	Average distance between connected points (*DCP*)	Y
	Average distance to the nearest neighbor (*DNN*)	Y
	Average distance between any two points (DAP)	Y
Similarity to the high-scoring essays' concept maps	Number of common nodes (*NCN*)	
	Number of common edges (*NCE*)	
	Edge similarity (*ES*)	

concept map, participants followed the rules similar to the automatic extracting process. To reduce the burden of concept recognition, some basic statistical information (word frequencies and a stop word list) from the text was provided. Participants were required to identify key concepts and the co-occurrence relationship between these concepts for each composition and generate a skeleton concept map (the weight of the concept edges was not included here). Two students independently constructed a concept map for each composition, and then the teacher led a discussion with the students about the differences between the two concept maps and determined the final manual concept map. Comparing the concept maps between the manually constructed and automatically extracted from the same composition allows us to assess the degree of consistency between them. We found that the participants recognized some concepts that were not considered by the automatic extraction process, so we optimized the extraction technique program, such as entering standardized word types for specific concepts. Finally, the number of identical nodes in the automatically constructed concept map and the manually constructed concept map is 93.8% on average; the number of identical edges in the automatically constructed concept map and the manually constructed concept map is 91.2% on average. Thus, the validity of the concept maps created from the essays should ultimately be reflected in the accuracy and effectiveness of the automatic essay scoring model in this study.

2.2 Features based on the concept maps

In this section, a graph-based feature set is computed based on a combination of a concept map structure and word embeddings representation. The assumption is that the ideas in an essay might change the structure and characteristics of a concept map extracted from the essay. Based on capturing the relationships between concepts, the distances between concepts are calculated based on the vector representation of concepts. We used the word2vec (Mikolov *et al.* 2013) representation, which is a well-known model used in natural language processing to represent words as a numeric vector. A numerical vector with 300 dimensions, by a skip-gram model, implies semantic information for each concept in this task, and the Euclidean distance metric was used to compute the distances between concepts. Based on the four characteristics of concept maps, 12 features were distilled and listed in Table 2.

https://doi.org/10.1017/S1351324921000103 Published online by Cambridge University Press

2.2.1 Global convergence

Convergence is one of the most important formal measurements when evaluating idea cohesion (Zhou, Luo and Chen 2018). Essays have one central object of elaboration—the main idea, which is unique and serves as the convergence point for all the information in the essay; the other ideas and content should be relevant to the main idea and extend outward with the main idea as the core. The main idea is the focus defined by the writer on the premise of their existing knowledge (Stevens, Slavin and Farnish 1991), and the selection of other subject information reflects the cognitive perspective of the writer on the main idea. Three proposed features to measure the global cohesion of ideas are as follows:

–Moran's I (*MI*)

Moran's I (MI) is a classic spatial autocorrelation measurement that expresses the global clustering situation of points in space. If the values of variables in space become more similar to the reduction of distance, it means that the data are clustered in space, which is called positive spatial correlation. If an article shows positive spatial autocorrelation, it indicates that the parts of the composition are well related to each other (Kim 2013). On the contrary, if the measured value grows with the reduction of distance, the data are scattered in space, which is called a negative spatial correlation. This indicates a lack of dependence on the composition, and the composition contains randomness. Zupanc and Bosnic (2017) method was adopted to calculate this measurement, which suits the researcher's 300-dimensional semantic space adjustment from the original two-dimensional space:

$$Moran's\, I = N/S \cdot n \sum_{k=1}^{n} \left[\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} \left(D_i^k - \overline{D_c^k}\right) \left(D_j^k - \overline{D_c^k}\right) \Big/ \sum_{i=1}^{N} \left(D_i^k - \overline{D_c^k}\right)^2 \right] \tag{1}$$

where N is the number of points (or concepts) in a concept map and n is the number of dimensions on the word2vec representation. So, $n = 300\ D_i^k\ k = 1 \ldots$ n; $i = 1, \ldots$, and N is a kth dimension of concept *i*; $\overline{D_c^k}$ is the kth dimension of a mean center. Weights (w_{ij}) are assigned to every pair of concepts, with the value $w_{ij}= 1$. If *i* and *j* are neighbors, then it means there is an edge between *i* and *j* ; otherwise, the value $w_{ij}= 0$ and S are a sum of w_{ij}. The range of *MI* varies from -1 to $+1$. A positive value indicates positive spatial autocorrelation, and the neighboring concepts cluster together, while a negative value indicates the opposite. Furthermore, values close to zero indicate complete randomness.

–Standard distance (*SD*)

Standard distance (SD) measures the amount of absolute dispersion in a concept map, and in this study, it is used to detect deviating concepts in essays by using a formula similar to standard deviation, this feature was also used by Zupanc and Bosnic (2017).

$$S_D = \sqrt{\sum\nolimits_{k=1}^{n} \sum\nolimits_{i=0}^{N} \left(D_i^k - \overline{D_c^k}\right)^2 / N} \tag{2}$$

N, n, D_i, and $\overline{D_c^k}$ in Equation (2) have the same meaning as Equation (1) because it is strongly influenced by extreme values like SD; therefore, atypical concepts have a dominant impact on the value of this feature.

–Mean of PageRank value (*MPR*)

PageRank (PR) (Brin and Page 1998) is a powerful ranking algorithm that is applied to aggregate website link choices. In a concept map, for the *i* node, the more nodes are linked to *i* , and the larger the PR value of the node linked to *i*, the larger the PR value of node *i* will be. Given a concept map structure, concepts with closer probabilities to the main idea tend to be more visited.

https://doi.org/10.1017/S1351324921000103 Published online by Cambridge University Press

That is, a concept that is related to the main idea in the essay tends to have very high connectivity, which means a high PR value, and an aggregate concept pattern close to the main idea will form.

According to the algorithm of PR, its value is also influenced by a number of concepts in the concept map. A concept map with more concepts will obtain a smaller PR value than a concept map with fewer concepts, even if the two concepts have the same linking structure. For this reason, the original PR value is multiplied by the total number to produce a normalized PR value (Somasundaran *et al.* 2016). Besides, because normalized PR values tend to be small numbers, the researchers made negative log versions for all normalized PR values (hereafter referred to as PR value). Finally, a feature corresponding to the mean PR value was computed for a concept map.

2.2.2 Local convergence

The degree of local convergence is mirrored by the degree of correlation among the different parts of a concept map. For example, the concept map for an essay with a clear but not well-supported main idea may show more segmentation and be divided into several subgraphs that are not related to each other or have low correlation. In this study, Gettis's G maximum number of clusters and the transitivity of the concept map are adopted to describe the degree of local cohesion.

–Local Gettis's G (*LG*)

In this study, the local Gettis's G is used to examine idea cohesion at a local scale. Global autocorrelation statistics may obscure the fact that local spatial autocorrelation is negative in space. Thus, analyzing whether some positions in the concept map are negatively correlated in space is interesting, because concepts with negative spatial autocorrelation may be weakly related to the main idea of the whole composition. Similar to Zupanc and Bosnic (2017), the researchers calculated Gettis's G for a 300-dimensional semantic space, where N, n, k, D_i^k, and w_{ij} have the same meaning as in Equation (1), and the larger the value of Gettis's G is, the greater the likelihood of local aggregation:

$$Gettis's\ G = 1/n \sum_{k=1}^{n} \left[\sum_{i}^{N} \sum_{j=1}^{N} w_{ij} D_i^k D_j^k \Big/ \sum_{i=1}^{N} \sum_{j=1}^{N} D_j^k D_j^k \right] \tag{3}$$

–Number of maximal cliques (*NMC*)

In graph theory, a clique is a set of vertices with edges between them. In a graph, if a clique is not contained by any other clique, then it is not a true subset of any other clique, and it is a maximal clique on the graph. In a concept map, concepts in a maximal clique tend to point to common ideas, and these concepts have common characteristics. At the same time, the concepts in a maximal clique are more or less connected with other concepts in the clique. Each maximal clique is a relatively aggregated unit on the graph, and the connections between the different maximal cliques are relatively loose.

–Graph transitivity (*GT*)

Transitivity is the main concept for many relational structures. In a graph, the calculation method is the fraction of all possible triangles which are in fact triangles, and possible triangles are identified by the number of "triads" that have two edges with a common vertex. Transitivity measures the probability that the adjacent vertices of a vertex are connected. Nafa *et al.* (2016) proposed that the semantic graph transitivity (GT) can be used to discover hidden cognitive relationships among knowledge units. Two concepts that connect to the same concept are likely to connect, the stronger the transitivity of the concept map, the closer the relationship between the pairs of concepts.

https://doi.org/10.1017/S1351324921000103 Published online by Cambridge University Press

2.2.3 Distance between nodes

The changes and developments in the concepts or exemplifications presented in compositions reflect the compactness coherence between contexts. Zupanc and Bosnic (2017) use similar metrics based on transforming sequential essay parts to access the coherence of essays.

–Average distance between connected points (*DCP*)

Connected points refer to the pair of points with an edge between them. For each concept map, the sum of all the Standard Euclidean distances are measured based on word2vec between two connected concepts and then dividing it by the number of connected concept pairs.

–Average distance to the nearest neighbor (*DNN*)

For a particular concept, the average distance to the nearest neighbor (*DNN*) is the average distance from each concept to its nearest connected concept. The Standard Euclidean distances are measured based on word2vec between a concept and its nearest neighbor.

–Average distance between any two points (*DAP*)

The average distance between any two points (*DAP*) is another metric to measure how well and fast ideas are developed in a composition. This number is represented as the average Standard Euclidean distance between all paired nodes in a concept map.

In addition, three reference features were computed only for accessing the value of the word embeddings in terms of system performance. We replaced the measures based on the Standard Euclidean distance between word2vec vectors with the word-based distance measures (Wu and Palmer 1994) using Cilin. This is the corresponding formula:

$$sim_{LC}\left(c_1, c_2\right) = \left(2 \times depth\left(lso(c_1, c_2)\right)\right) \Big/ \left[len\left(c_1, c_2\right) + 2 \times depth\left(lso\left(c_1, c_2\right)\right)\right] \tag{4}$$

Equation (4) $depth\ (c_i)$ indicates the depth of the concept c_i in Cilin, and $len\ (c_1, c_2)$ is the shortest distance between the concepts c$_1$ and c_2 (calculated by the number of edges), and $lso\ (c_1, c_2)$ refers to the deepest common parent node of c$_1$ and c_2 in Cilin. Then, three reference features (denoted as DCP', DNN', and DAP'), which correspond to the above three features (*DCP, DNN,* and *DAP*), are generated.

2.2.4 Similarity to the high-scoring essays' concept maps

After drawing lessons from the Chinese composition similarity analysis method of LSA (Cao and Yang 2007) (based on the edge matching method in graphs), the matching degree between concept graph structures is calculated to represent the similarity between texts in this study.

In this study, the high-scoring compositions refer to the highest-scoring essays written by students for each prompt in the training set (i.e., idea score of 6 in prompts 1, 2, and 3, and score of 4 in prompts 4, 5, and 6), and each prompt gets a subset of high-scoring essays. Each subset of high-scoring compositions is mapped into a large concept map via a similar method discussed in Section 2.1. Then, the concept map for each target essay in a prompt is compared with the same large concept map for this prompt to calculate the similarity between the target essay and the high-scoring compositions. It is assumed that the more the edge structures of the two concept maps repeat, the greater the similarity. Lastly, the information (nodes, edges, and weights) drawn from the target essay and the high-scoring essay subsets are used to calculate the similarity.

–Number of common nodes (*NCN*)

The number of common nodes (*NCN*) is the number of identical nodes in the concept map of a particular composition and the concept map of the high-scoring composition from the same prompt.

https://doi.org/10.1017/S1351324921000103 Published online by Cambridge University Press

–Number of common edges (*NCE*)

The number of common edges (*NCE*) is the number of identical edges in the concept map of a particular composition, and the concept map of a corresponding high-scoring composition.

–Edge similarity (*ES*)

The edge similarity (*ES*) calculation formula used in this study is as follows:

$$\text{Similarity}\,(CM_i, CM_H) = m_c/\min\,(CM_I, CM_H) \tag{5}$$

In Equation (5), the similarity between a concept map (CM_i), and a concept map from a high-scoring essay set (CM_H) is noted as similarity (CM_i, CM_H), which is taken as the similarity between an essay (e_i) and high-scoring essays. m_c is the number of the identical edges of CM_i and CM_H, and min (CM_i, CM_H) is the minimum number of edges of CM_i and CM_H.

3. Implementation and evaluation

3.1 Data

Experiments were performed on six different prompts (essay topics) using six different datasets. The essays used in this experiment are from ATQE 2016 writing tasks and were written by eighth-grade middle school students. The writing task required students to write a composition based on a specific topic. Students have to focus on expressing the main idea and supporting it with exemplifications and details. The datasets in this study contain student essays from six different prompts; the first three prompts included different writing genres (persuasive, expository, or narrative). The topics of these three prompts in the sequence are “Companionship is the best gift,” “If given another chance,” and “Something hard to forget.” The last three prompts (4, 5, and 6) are practical writing: prompt 4 is a notification for an event, prompt 5 is a proposal to save food, and prompt 6 is a letter to your teacher. Appendix A shows the translation of prompt 1 into English.

Human raters of large-scale examinations usually need to have certain qualifications. The essay raters that participated in this study were mainly composed of teachers, teaching researchers, and graduate students. All the raters have experience with scoring and received face-to-face on-site training before scoring essays for this research project. The training mainly included scoring rules, composition examples, scoring practice, and final assessments. The assessment samples used in training constituted typical student responses. These essays were given benchmark scores by the expert committee in advance. If a rater’s score differed greatly from the benchmark score, he or she did not pass the assessment and needed to be retrained. The raters who passed the assessment are qualified to formally score compositions independently. The ideas of each essay in the examination were rated by two human raters on an integer scale of 1 to 6 (poor (1) to excellent (6)) for prompts 1, 2, and 3, and on an integer scale of 1 to 4 (poor (1) to excellent (4)) for prompts 4, 5, and 6. If there was a notable discrepancy between scores, the expert raters arbitrated the discrepancy. Compositions needed to be arbitrated by experts on average of 2–3% of the time. It is noted that the score after arbitration was used as the reporting score for the test rather than used in this study. Inter-rater reliability was calculated using quadratic weighted kappa (QWK) (Cohen 1968). QWK for the six prompts are as follows: 0.75, 0.71, 0.74, 0.79, 0.76, and 0.80, which is considered acceptable.

Each dataset was divided into training and test sets, and the ratio of the training set to the test set was 7:3. The training set was used to develop the scoring model, and the test set was utilized to predict the scores and evaluate the prediction accuracy. Table 3 reports the characteristics of each dataset.

https://doi.org/10.1017/S1351324921000103 Published online by Cambridge University Press

Table 3. Description of the training set and test set

		Prompt					
		1	2	3	4	5	6
Training set	Number of essays	2100	2100	2100	1400	1400	1400
	Mean number of characters	433.60	426.20	414.10	158.47	129.00	154.36
	SD of number of characters	153.38	150.22	139.20	47.03	47.35	51.11
	Range of the rubric	0-6	0-6	0-6	0-4	0-4	0-4
	Mean of human score	4.06	4.08	3.67	2.91	2.69	3.07
	SD of human score	1.15	1.22	0.86	0.50	0.70	0.60
Test set	Number of essays	900	900	900	600	600	600
	Mean number of characters	432.58	428.50	415.84	157.11	128.61	153.84
	SD of number of characters	154.42	149.97	133.74	47.75	49.34	51.97
	Mean human score	4.08	4.05	3.70	2.90	2.65	3.08
	SD of human score	1.12	1.21	0.86	0.52	0.77	0.65

SD = standard deviation.

3.2 Prediction model

A multiple linear regression (MLR) model was adopted to predict the idea scores in this study for three reasons:

(1) The first human scores (hereafter referred to as H1) available for each essay were used to train the MLR model. The MLR generated a continuous value prediction for the score. If a prediction was outside the range of the scale (such as a predicted score above 6 or less than 1), it was assigned the value of the nearest score point. Scores predicted by MLR can be used to assess how much a student has progressed in their writing ability. The kappa measurements of agreement (introduced in Section 3.4) were calculated between the MLR predicted scores, which were rounded to the nearest integer and the human scores.

(2) The researchers sought to use meaningful features to evaluate the ideas of an essay instead of dealing with a large number of variables that are more difficult to interpret, and MLR is sufficient to handle a relatively small number of variables.

(3) With MLR, the researchers were able to explore the importance of features by examining their relative weights in the model. Through this study, it is hoped to construct a relatively stable framework and weighting scheme of features for different writing tasks. Hence, we investigated whether the prediction accuracy of the regression models trained on the different prompt sets is stable across six datasets and assess the relative importance of features, which is useful to improve the standardization, interpretability, and generalizability of AEE.

After considering these reasons, a multiple regression model was used to predict the ideas score.

3.3 Baseline

3.3.1 Fine-tuning BERT

This study used two baselines, the bidirectional transformer architecture (BERT) model, which is pretrained and fine-tuned on our data, and the Coh-Metrix. They served as the baselines to compare the performance of the concept map-based features.

The field of NLP research has been dominated by deep neural network research, such as the BERT model, which has proven to have great advantages in other fields. BERT is a pretrained language model that has been proven effective in various NLP tasks. The use of neural models for automatic essay scoring is still in the initial exploratory stage (Rodriguez, Jafari and Ormerod 2019; Mayfield and Black 2020). Neural models use a large amount of existing text data to pre-train multi-layer neural networks with context-sensitive meaning. They also contain more than 100 million parameters that are then fine-tuned to a specific new labeled dataset and used for classification or prediction.

In our work, we use training data to fine-tune a pretrained BERT model (Devlin *et al.* 2019) that produced a 768-dimensional embedding based on a network.[e] For texts, BERT tokenizes them using predefined vocabulary and then generates the input sequence by concatenating a [CLS] token, tokenized text and a [SEP] token. Generally, the hidden state corresponding to the [CLS] token is used as the text representation of the input text (Devlin *et al.* 2019). Here we have to truncate a small number of essays longer than 512 Chinese characters since it is the maximum the BERT can process, which is mostly in datasets 1, 2, and 3. For the regression task in this study, a specific linear layer is used on top of BERT to obtain a value prediction, and BERT is fine-tuned together with the linear layer by the Adam optimizer, the set learning rate is 0.00001. In addition, we train the network by concatenating features derived from the concept maps with the text representation produced by BERT to compare the performance of the BERT model without the proposed features. The technical contribution of these baselines is to investigate the performance of the fine-tuned BERT model with our training data and whether a fine-tuned BERT model would offer different performances with and without using concept-map-based features.

3.3.2 Coh-Metrix features

Coh-Metrix, which was first used for text analysis of reading comprehension, is an online English text analysis system developed by the University of Memphis. Coh-Metrix has been widely used in various fields of research and represents the current mature level of text mining (Graesser *et al.* 2004; Zedelius, Mills and Schooler 2018) and automated essay scoring (Aryadoust and Liu 2015; Mo 2018; Latifi and Gierl 2020; Shin and Gierl 2020). In this study, a Chinese Coh-Metrix feature system, inspired by English Coh-Metrix, was developed. It contains 68 features under 8 dimensions of language characteristics. Appendix B has an explanation of the computational features for Chinese. This study investigated how a feature set drawn from concept maps performs in comparison to the Coh-Metrix baseline.

3.4 Evaluation

The evaluation of AEE includes the prediction accuracy and quality analysis of AECC-I. Since an important goal of AEE is to imitate human scoring, the more accurate the model, the closer the two scores will be. For accessing the accuracy of the proposed AECC-I system, the following statistical measurements were used.

3.4.1 Quadratic weighted kappa (QWK)

In 1960, Cohen proposed that the kappa value can be used as an indicator for rating the consistency of graders and has been widely used in practice ever since. The kappa value ranges from 0

[e] See https://github.com/google-research/bert for detail.

https://doi.org/10.1017/S1351324921000103 Published online by Cambridge University Press

(random agreement between raters) to 1 (complete agreement between raters). In general, a kappa value of 0.7 is an acceptable minimum in essay scoring (Ramineni and Williamson 2013) since it explains nearly half of the scoring variation. The QWK takes into account the degree of difference between different scores (Fleiss and Cohen 1973). For example, the difference between 1 point and 4 points is much larger than that between 1 and 2 points. In this study, the QWK was computed to investigate the agreement between automated scoring and human scoring. QWK is given by:

$$\mathrm{QWK} = 1 - \sum_{i,j} w_{i,j} O_{i,j} / \sum_{i,j} w_{i,j} E_{i,j} \tag{6}$$

w are weights calculated as:

$$w_{i,j} = \left(i - j\right)^2 / (r - 1)^2 \tag{7}$$

where O is the confusion matrix of observed H1 counts and predicted scores and E is the confusion matrix of expected H1 counts and predicted scores based on chance. The index i and j refer to a rating of the *H1* i score point and a score point j by the AEE. The variable r is the number of rating categories.

–Adjacent agreement measure

This measurement is defined as the percentage of essays that were scored equally and similarly (usually within one point) by human raters and the AEE system.

To investigate the internal structure of the feature set, a correlation analysis was applied (Subsection 4.2.1) to investigate the relation between features and scores, and applied factor analysis (Subsection 4.2.2) was used to explore the latent components of construct measured by these features.

4. Results and discussion

4.1 Concept map-based AEE system accuracy

The following three different models were compared to evaluate if concept-map-based features could yield better model performance for automated scoring ideas:

(1) Automated scoring for the ideas using a fine-tuned BERT model without concept-map-based features. The baseline is denoted as M_B;

(2) Automated scoring for the ideas using a fine-tuned BERT model that combines concept-map-based features. The baseline is denoted as M_{B+CM};

(3) Automated scoring for ideas using only Coh-Metrics features (listed in Appendix B, Table E.1). The baseline features are denoted as M_{Coh};

(4) Automated scoring of the ideas used features from concept maps (listed in Table 2), and this scoring model is denoted as M_{CM}. Furthermore, the models corresponding to the four feature subsets under M_{CM} are denoted as M_{CM-c}(features of global cohesion of ideas: *MI*, *SD*, and *MPR*), M_{CM-l} (features of local cohesion of ideas: *LG, NMC, and GT*), M_{CM-d} (features of idea development: *DCP, DNN, and DAP*), and M_{CM-s} (features of similarity to high-scoring essays);

(5) To assess the value of the embedding vector in terms of system performance, the features that do not rely on word embeddings in Table 2 are used to build a model. They are denoted as $M_{CM'}$ and are compared to M_{CM}, which does rely on embeddings. Accordingly, $M_{CM'-c}$ is built on the features of *MI* and *SD*; $M_{CM'-l}$ is built on the features of *NMC* and *GT*. Since the feature group of "Distance between nodes" is all based on word embeddings, $M_{CM'-d}$ is built on reference features of *DCP*′, *DNN*′,and *DAP*′(described in Section 2.2). These

https://doi.org/10.1017/S1351324921000103 Published online by Cambridge University Press

features use the word distance measurements based on Cilin rather than word embeddings, while M_{CM-s}, which does not use distance measurements, is preserved.

(6) Both concept map features M_{CM} and Coh-Metrics were used for scoring ideas, and the scoring model for ideas is denoted as M_{CM+}.

Table 4 reports the R^2 and QWK values of the various models for ideas scoring. R^2 is the fraction of the variance in the data that are explained by the regression, and the values range from 0 to 1. The closer the value is to 1, the stronger the explanatory power and the better the model fit. The results show that all QWK values from the M_{CM} model are in the range of 0.77 to 0.90 on the six datasets.

The M_{CM} model outperformed the Coh-Metrics baseline in all datasets, and the QWK values were about 8–10% higher than the QWK values of M_{Coh}. When the concept map features were added to the Coh-Metrics features, there was a prominent boost in performance (more than 10%). Relatively, the Coh-Metrics baseline did not substantially improve the performance of M_{CM+}.

It is noted that both BERT models (M_B) achieve approximately the same performance as the Coh-Metrix feature-based model; however, our data show that the BERT models did not perform better than the concept map feature-based models on most prompts. Even the models combined with the concept map-based features(M_{B+CM}) only exceeded M_{CM} by 0.01 QWK on prompt 3. It is generally believed that the relative performance of a fine-tuned BERT model depends on the similarity between the pretraining task and the target task. So, a possible explanation for this result is that the knowledge learned by our BERT models has a limited ability to access the quality of ideas in essays. For long essays (prompts 1, 2 and 3), the BERT models leave out some words, which may cause accuracy problems because some of the semantic information is lost; for short essays, the semantic differences among essays are relatively small, but the global organization and structure of the concepts may show distinctions.

We also noticed a result that was not expected. A simple combination of proposed features and the BERT model presented small improvements of 0.02–0.05 QWK compared to the BERT models without additional features. Here, we do not claim that our results are the best results that BERT fine-tuning can achieve. A more sophisticated combination would likely yield better results via optimization since the complexity of hyperparameter and curriculum learning for transformers could be improved on larger training datasets.

Table 4 also reports the performance of the four smaller feature set models. The QWK values of M_{CM-c} are the best-performing individual small feature set and is basically equal to that of M_{Coh}, while the performance of the other three models are ranked from highest (*1^st^*) to lowest (*3^rd^*): 1st. M_{CM-l}, *2nd.* M_{CM-d}, and 3*rd*. M_{CM-s}. This confirms the hypothesis that the feature sets for the global cohesion of ideas give the best performance for scoring ideas.

In order to assess the value of the word embeddings used in the features extracted, the performance of $M_{CM'}$ is compared with M_{CM}. The QWK values of M_{CM} are all higher than $M_{CM'}$ in a range of 6–12%. The feature group based on embeddings used in M_{CM-d} has the greatest improvement in terms of model performance. Compared with the Cilin-based word distance features in $M_{CM'-d}$, the embeddings-based features in M_{CM-d} have approximately doubled the performance of $M_{CM'-d}$. While the PR feature in the global convergence feature group performs best, after the other two features based on embeddings were added to the model, the QWK values are improved by 5–9 points on the six datasets. In the local convergence feature group, the only feature-based on embeddings (*LG*) increased the KAPPA value of the model based on the two features (*NMC, GT*) by 3–7%.

In this study, the exact agreements and the exact-plus-adjacent agreements were calculated. Table 5 shows that the exact agreements for M_{CM} and M_{CM+} are all higher than the human raters. The distribution of M_{CM} is closer to the human scores than the baseline model, while M_{CM+} is the closest to the human scores. M_{CM} improves the exact agreements with H1 by 3–8% compared to M_{Coh}. To compare the differences between models, McNemar's test was used on the exact and

https://doi.org/10.1017/S1351324921000103 Published online by Cambridge University Press

Table 4. The R^2 for multiple linear regression and QWK value (Quadratic-weighted kappa between automated and human scoring) for the feature sets for ideas scoring on the test set

		Prompt					
System		1	2	3	4	5	6
M_B	R^2	0.61	0.55	0.60	0.61	0.67	0.64
	QWK	0.76	0.70	0.77	0.73	0.82	0.79
$[M_{B+CM}]$	R^2	0.64	0.59	0.63	0.69	0.75	0.68
	QWK	0.79	0.73	0.79	0.78	0.85	0.82
M_{Coh}	R^2	0.65	0.61	0.62	0.72	0.68	0.73
	QWK	0.76	0.70	0.69	0.80	0.78	0.81
$M_{CM'}$	R^2	0.68	0.63	0.60	0.70	0.61	0.65
	QWK	0.79	0.71	0.66	0.78	0.75	0.79
$M_{CM'-c}$	R^2	0.57	0.52	0.54	0.56	0.50	0.55
	QWK	0.71	0.64	0.63	0.70	0.66	0.71
$M_{CM'-l}$	R^2	0.55	0.51	0.48	0.50	0.48	0.49
	QWK	0.66	0.58	0.57	0.64	0.60	0.59
$M_{CM'-d}$	R^2	0.23	0.20	0.18	0.21	0.19	0.21
	QWK	0.30	0.27	0.25	0.32	0.28	0.34
$M_{CM'-s}$	R^2	0.60	0.59	0.48	0.52	0.47	0.53
	QWK	0.68	0.70	0.53	0.70	0.57	0.69
M_{CM}	R^2	0.75	0.70	0.69	0.79	0.78	0.80
	QWK	0.85	0.80	0.78	0.88	0.87	0.90
M_{CM-c}	R^2	0.62	0.59	0.61	0.62	0.59	0.63
	QWK	0.76	0.72	0.70	0.78	0.71	0.80
M_{CM-l}	R^2	0.59	0.56	0.54	0.60	0.55	0.53
	QWK	0.69	0.61	0.62	0.72	0.67	0.64
M_{CM-d}	R^2	0.49	0.41	0.49	0.51	0.37	0.52
	QWK	0.59	0.54	0.58	0.63	0.52	0.65
M_{CM-s}	R^2	0.60	0.59	0.48	0.52	0.47	0.53
	QWK	0.68	0.70	0.53	0.70	0.57	0.69
M_{CM+}	R^2	0.77	0.72	0.72	0.79	0.78	0.82
	QWK	0.87	0.83	0.80	0.90	0.89	0.91
p-value	$M_{Coh} - M_{CM}$	0.00[a]	0.00[a]	0.00[a]	0.00[a]	0.00[a]	0.00[a]
	$M_B - M_{CM}$	0.00[a]	0.00[a]	0.01[a]	0.00[a]	0.00[a]	0.00[a]
	$M_{B+CM} - M_{CM}$	0.00[a]	0.00[a]	0.02[a]	0.00[a]	0.00[a]	0.00[a]

[a] *p*-value 0.05.

Table 5. The exact (*E*) agreements and exact-plus-adjacent (*E*+*A*) agreements for the six datasets

Agreement	Prompt 1		Prompt 2		Prompt 3		Prompt 4		Prompt 5		Prompt 6	
	E	$E+A$	E	$E+A$	E	$E+A$	E	$E+A$	E	$E+A$	E	$E+A$
$H1-H2$	0.63	0.89	0.58	0.87	0.60	0.81	0.74	0.90	0.70	0.87	0.78	0.92
$H1-M_{Coh}$	0.67	0.86	0.60	0.85	0.57	0.80	0.77	0.91	0.73	0.86	0.79	0.94
$H1-M_B$	0.67	0.87	0.60	0.84	0.66	0.85	0.70	0.84	0.79	0.90	0.76	0.90
$H1-M_{B+CM}$	0.69	0.90	0.64	0.87	0.68	0.89	0.77	0.90	0.80	0.92	0.77	0.91
$H1-M_{CM'}$	0.68	0.89	0.65	0.87	0.60	0.84	0.79	0.95	0.76	0.88	0.82	0.95
$H1-M_{CM}$	0.70	0.91	0.67	0.90	0.67	0.86	0.84	0.97	0.80	0.91	0.87	0.99
$H1-M_{CM+}$	0.72	0.96	0.69	0.91	0.68	0.87	0.88	0.98	0.83	0.94	0.89	0.99
						p-value						
$M_{Coh}-M_{CM}$	0.01[a]	0.00[a]	0.00[a]	0.00[a]	0.00[a]	0.00[a]	0.00[a]	0.00[a]	0.00[a]	0.00[a]	0.00[a]	0.00[a]
M_B-M_{CM}	0.01[a]	0.00[a]	0.00[a]	0.00[a]	0.02[a]	0.03[a]	0.00[a]	0.00[a]	0.04[a]	0.03[a]	0.00[a]	0.00[a]
$M_{B+CM}-M_{CM}$	0.03[a]	0.03[a]	0.01[a]	0.01[a]	0.03[a]	0.01[a]	0.00[a]	0.00[a]	0.05	0.03[a]	0.00[a]	0.00[a]

[a]*p*-value 0.05.

exact-plus-adjacent agreements. The results show significant differences in the exact agreements between models M_{Coh} and M_{CM} for predicting idea scores.

Figure 3 shows that the distribution patterns of the automated scoring models are roughly the same as human scores. The scores distribution of M_{CM} is closer to the human scores. The following studies were conducted using M_{CM}, and the system AECC-I is based on this model.

4.2 Evaluation of the proposed features

An end-to-end AEE system can achieve high consistency with manual scoring in many cases, but these models do not address the structural validity (Perelman 2012; Condon 2013; Rodriguez, Jafari and Ormerod 2019). AECC-I is based on a rubric's idea scoring for compositions to ensure that the features used for scoring can cover the dimension of the construct. This allows the feature data inside the AEE to convert into formative feedback information (Zhang *et al.* 2019) and help students understand why their scores are low.

4.2.1 Correlation analysis

The construct coverage of AEE depends on the internal structure and must have feature sets with a proven ability to evaluate ideas. Applied correlation analysis was used to investigate the relationship between each concept map feature in relation to the idea scores. The following analysis is mainly based on MLR models, and the Pearson correlation coefficient between each feature and ideas score are shown in Table 6. Besides, the composition length tends to affect their scores, as does the scoring of ideas because longer essays are likely to get higher scores. Therefore, the influence of the composition length as a variable was removed, and the partial correlation between other features and scores was calculated. Due to space limitations, only the evaluation of features in prompt 1 and prompt 2 were performed.

All the features have high (> 0.5) to low (between 0.1 and 0.3) correlation with the idea scores (significant at $p < 0.01$). The three global cohesion ideas features are all moderately or highly positively correlated with ideas ($p < 0.01$). This indicates that the stronger the global cohesion is, the

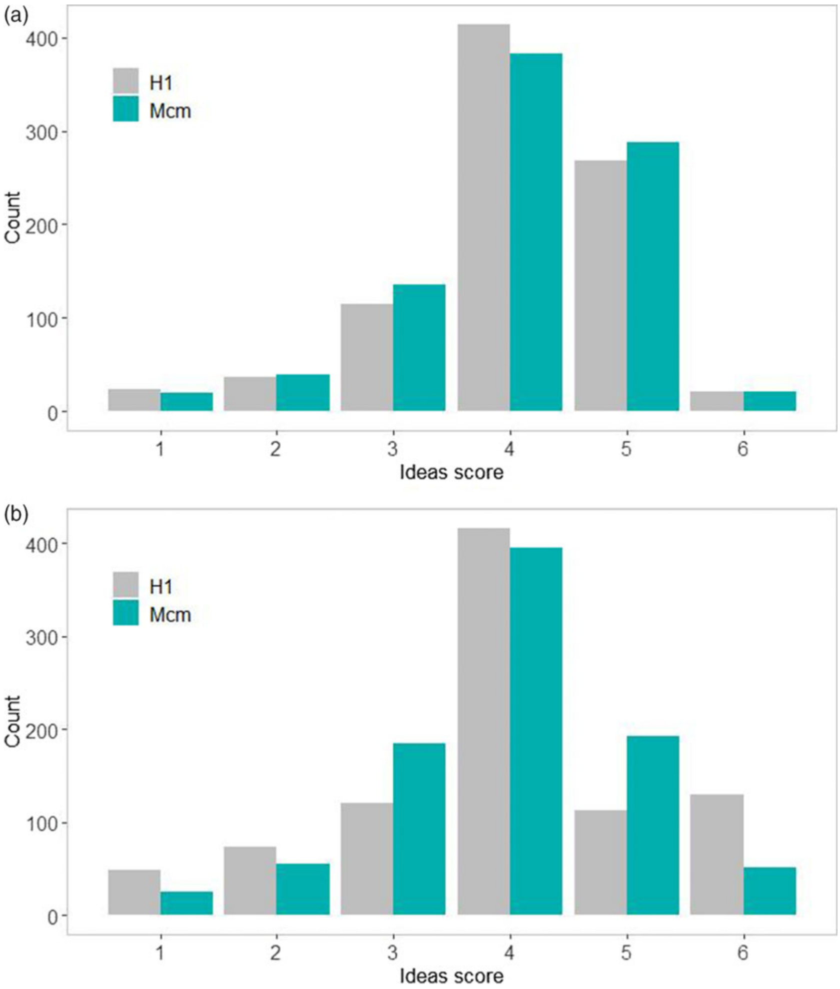

Figure 3. Comparison of the score distributions between automated and human scoring on two prompts.

higher the idea scores will be. This is consistent with previous research results (Li *et al*. 2012; Ke, Zeng and Luo 2016; Antiqueira *et al*. 2017). The local Gettis's G and the number of maximal cliques (NMC) have a negative correlation with the scores, and this means that the more local aggregation in a concept map, the more cliques a concept map will be divided into, and the lower the ideas scores of the composition will be. The GT has a positive correlation with the idea scores, and this suggests better connectivity in a concept map and higher idea scores. All of the idea development features show a positive correlation with scores, and this indicates that the larger the change from point to point in the concept map of composition, the more adequate the development of the ideas in the composition will be, which is consistent with Ke, Zeng and Luo (2016). The three features of similarity in relation to the high-scoring essays all have a positive correlation with the idea scores and are consistent with the researcher's expectation that the more similar a concept map of a composition is to a high-scoring composition, the higher the ideas score will be. After the composition length effects were removed, all the concept map features affected by composition length had a reduction in correlation with the ideas score (see the Partial Correlation column). Furthermore, no significant correlation with the idea scores in relation to the average distance between connected points, the average distance between any two points on prompt 1, the NMC, and the average distance to the nearest neighbor.

Table 6. Correlation and partial correlation (controlling for length) of concept map features with idea scores

	Prompt 1		Prompt 2	
Feature	Corr.	Partial corr.	Corr.	Partial corr.
Moran's I	0.40**	0.22**	0.41**	0.18**
Standard distance	0.40**	0.34**	0.39**	0.15**
Mean of PageRank value	0.62**	0.54**	0.63**	0.20**
Local Gettis's G	−0.38**	−0.12**	−0.33**	−0.11**
Number of maximal cliques	−0.51**	−0.05∗	−0.44**	−0.02
Graph transitivity	0.42**	0.18**	0.34**	0.10**
Average distance between connected points	0.11**	0.02	0.11**	0.05*
Average distance to the nearest neighbor	0.18**	0.05*	0.11**	0.03
Average distance between any two points	0.17**	0.03	0.18**	0.06∗
Number of common nodes	0.61**	0.14**	0.59**	0.06∗
Number of common edges	0.36**	0.17**	0.35**	0.10**
Edge similarity	0.20**	0.12**	0.35**	0.16**

**significant at $p < 0.01$; *significant at $p < 0.05$.

We initially extracted a large number of concept map features; however, the meaning of some features is unclear, and there is a high correlation between some features, which is not conducive for training the MLR model. Therefore, in addition to ensuring the fitting degree of the model, we also considered the following aspects when selecting features: 1) the relationship between the meaning of the feature and the idea's construct is not contrary to common sense; 2) some features that are not related to the conceptual interpretation of ideas' quality have been eliminated; 3) some features which are theoretically reasonable but statistically limited are eliminated to try to prevent the problem of multicollinearity, and 4) factor analysis was used to verify the construct validity of the feature set.

4.2.2 Factor analysis

Factor analysis is an important method to validate the scoring construct (Attali 2007) by exploring the internal structure of the proposed features.

Table 7 presents the Promax-rotated principal factor pattern for the 12 features for the four-factor solution. The 12 features were grouped into different factors based on their highest loading on the different factors due to structural validity, which is also coincident with what we expected. To show the influence of these four factors on the quality of the ideas more intuitively and concisely, the four factors were named as the main idea factor (measured by the *MI*, mean of the PR value, and SD features), a local support factor (measured by the Local Gettis's G, the NMC and GT features), a development factor (measured by the average distance between connected points (*DCP*), average *DNN* and the average distance between any two point features), and a similarity factor (measured by the *NCN*, the *NCE* and *ES* features).

4.2.3 Relative importance of features

In order to display the relative importance of the different features in the regression prediction, the contribution of each feature prediction to R^2 was calculated. Figure 4 presents the relative

Table 7. Factor pattern after Promax rotation for the four factors on prompts 1 and 2 and the factor names are *F1 (main idea), F. (local support), F3 (idea development), and F4 (similarity)*

	Prompt 1				Prompt 2			
Features	F1	F2	F3	F4	F1	F2	F3	F4
Moran's I	0.89	0.33	0.08	0.35	0.89	0.51	0.45	0.19
Mean of PageRank value	0.9	0.46	0.23	0.67	0.96	0.5	0.34	0.53
Standard distance	0.85	0.26	0.69	0.29	0.93	0.33	0.52	0.15
Local Gettis's G	−0.21	−0.89	−0.24	0.12	−0.36	−0.9	−0.23	0.23
Number of maximal cliques	0.33	−0.89	0.13	0.28	0.53	−0.89	0.2	0.17
Graph transitivity	−0.16	0.82	0.08	0.01	−0.32	0.87	−0.11	0.05
Average distance between connected points	0.3	−0.13	0.91	0.02	0.25	0.05	0.93	0.09
Average distance to the nearest neighbor	−0.07	0.4	0.77	−0.4	−0.17	0.58	0.7	−0.43
Average distance between any two points	0.45	0.16	0.95	0.04	0.35	0.26	0.98	0.04
Number of common nodes	0.59	0.5	0.21	0.8	0.81	0.54	0.23	0.65
Number of common edges	0.07	−0.29	−0.23	0.87	0.41	−0.15	−0.04	0.98
Edge similarity	−0.71	−0.01	−0.49	0.89	0.41	−0.15	−0.04	0.98

importance of the features for the regression models, and the global cohesion feature set of ideas makes the largest contribution. Next, the feature set for local cohesion of ideas and similarity to high-scoring essays. Lastly, the feature set for idea development has the smallest contribution in datasets 1 and 2. For individual features, the mean of the PR value contains the most important features on both prompts. The importance of most features is relatively stable, except for the SD and the features related to similarity.

Table 8 presents the standardized coefficients and p-values of the regression analysis, and the ranking of the standardized regression coefficients of the features and the ranking of the relative weights are similar.

4.3 Qualitative analysis and smart feedback

4.3.1 Qualitative analysis

To investigate whether concept map features indeed capture the idea characteristics, this study drew on the method of qualitative analysis in Somasundaran *et al.* (2016). The contents of the high-scoring essays for prompt 1 (with the topic of "Companionship is the best gift") were modified in three ways to simulate different defects. These modification methods are used to reduce the quality of composition ideas, obtain lower feature scores, and examine whether the feature scores are sensitive to composition quality. These three modification strategies lead to the scoring change of the first three factors (main idea, local support, and ideas development) described in Subsection 4.2.2. The last similarity factor, which was obtained by comparing it with a high-score composition, is not the characteristics of the composition itself. Therefore, this similarity feature is not included in this part of the analysis. Listed below are the types of compositions generated by the three modification strategies:

(1) No main idea: The main idea in this simulated composition is not clear. We found sentences or paragraphs in the composition that are closely related to the central idea and

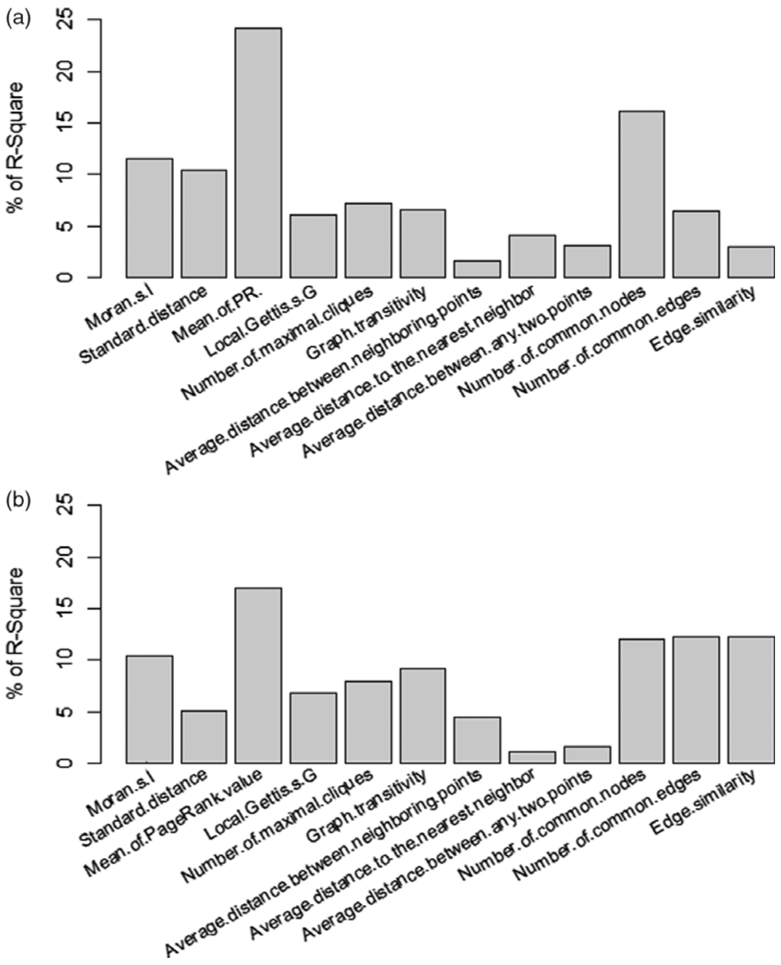

Figure 4. Relative feature importance is expressed as a percent of the total weights from regression for the predicted ideas scores.

then replaced all the other parts of the composition with this so that the full text contains similar ideas but has no focus. The length of this simulated essay changed a little between before and after the revision;

(2) Weak support: Various parts of this simulated essay lack support for the main idea. The ideas and examples in paragraphs two and three were deleted. Therefore, this essay is shorter than the original text.

(3) Vague development: This simulated composition presents slower and vaguer idea developments than the original text. To do this, some examples in paragraphs two and three were deleted and replaced with examples from the first paragraph. The length of this simulated essay is basically the same as the original text.

Table 9 presents three types of simulated composition information: the average score for the factor scores before and after, and composition length. The factor score is the weighted average of the standardized values of its associated features, and the weight is the load of the features on the factors in the factor analysis. Model M_{CM} was used to predict the idea scores for these simulated compositions (last column in Table 9). The impact of the three factors on the idea scoring ranked

Table 8. The regression table for the two prompts with the standardized coefficients and *p*-values

	Prompt 1		Prompt 2	
Features	Standardized coefficients	*p*-Value	Standardized coefficients	*p*-Value
Moran's I	0.50	0.00[a]	0.39	0.00[a]
Standard distance	0.46	0.00[a]	0.27	0.00[a]
Mean of PageRank value	1.67	0.00[a]	0.95	0.00[a]
Local Gettis's G	−0.42	0.00[a]	−0.30	0.00[a]
Number of maximal cliques	−0.45	0.00[a]	−0.27	0.00[a]
Graph transitivity	0.44	0.00[a]	0.36	0.00[a]
Average distance between connected points	0.13	0.03[b]	0.26	0.00[a]
Average distance to the nearest neighbor	0.31	0.00[a]	0.10	0.04[b]
Average distance between any two points	0.28	0.00[a]	0.12	0.02[b]
Number of common nodes	0.72	0.00[a]	0.45	0.00[a]
Number of common edges	0.43	0.00[a]	0.51	0.00[a]
Edge similarity	0.28	0.00[a]	0.51	0.00[a]

[a] *p*-Value < 0.01.
[b] *p*-Value < 0.05.

Table 9. Factor scores for simulated compositions

		Factor scores				
Composition	Text length	Main idea	Local support	Ideas development	Similarity	Predicted value Idea score
Original	588	1.75	0.13	0.20	1.44	6.00
No main idea	435	0.08	0.12	0.02	0.48	3.12
Original	538	0.36	0.13	0.37	1.15	6.00
Weak support	272	0.86	0.27	0.16	0.38	4.24
Original	554	0.32	0.15	0.19	1.09	6.00
Vague development	543	0.40	0.13	0.13	0.67	4.41

Text length = the number of Chinese characters.

as follows from highest to lowest: the main idea, local support, and ideas development. Therefore, the composition scores with different revision strategies were ranked from lowest to highest: no main idea, weak support, and then vague development.

Figure 5 intuitively illustrates whether the factor scores are sensitive to changes in the quality of the composition ideas. The same letters are used to represent each original essay, and its transformed version so that the difference in factor scores between the original and their transformed version is demonstrated in the graph. Compared to the original composition, the main idea score for the simulated composition is much lower than that of the original composition for most essays in Figure 5(a). It is interesting that even though the length of the weak supporting composition is shorter than that of the original text, the focus of the composition seems to be clearer than the original text. This indicates that deleting some of the examples in the text made the main idea

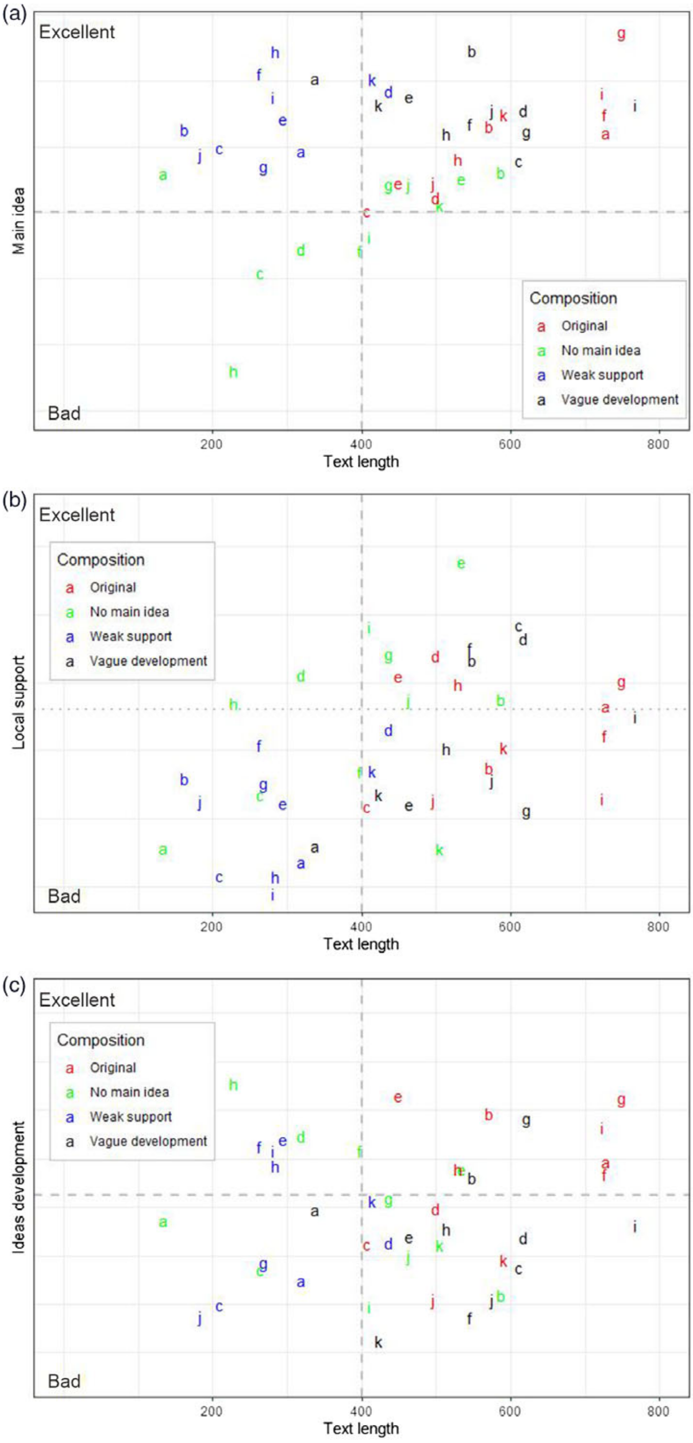

Figure 5. The changes of three-factor scores in different types of simulated composition.

https://doi.org/10.1017/S1351324921000103 Published online by Cambridge University Press

of the composition better. In Figure 5(b), the score of the weak supporting composition is lower than the score of the original text, while the score of the other two compositions did not change much. For the composition with poor development, the average score of development in Table 9 decreased, but the change in the scatter graph (c) was not obvious. The similarity factor scores are also shown in the table, and they indicate that the manipulations used to produce essays with no main idea, weak support, and vague development also made the resulting essays less similar to the original ones.

This result illustrates that the main idea is the most important factor for ideas scoring, and it is consistent with the factor analysis and satisfies the expectations for idea scoring constructs in Chinese compositions. Moreover, although weak support is the lowest one, the predicted scores are higher than the essay's labeled no main idea, which is due to the fact that it has been stripped of exemplification information, which strengthens the focus of the text, and it indicates that the concept map features are not easily affected by the text length.

4.3.2 Visual inspection and smart feedback

An inspection of the concept maps (See Figure 6) revealed a clear distinction between poor and excellent ideas in different compositions.

Figure 6(a) depicts a clear and meaningful center in the concept map with excellent ideas because the center of the concept map consists of three key concepts "myself," "thanks," and "parents." Other concept centers connect to the main idea and support it. Similar to the reference essay from Kim's research (2013), this concept map is highly coherent and connected. Although the word "company" does not appear in the text, the main idea is clearly expressed. From the concept map, it can be inferred that the author is thanking his parents for accompanying him in his growth. This composition will get a high ideas score, according to the rubric of ATQE, because the composition has a well-focused main idea and expresses ideas well and effectively addresses the topic. It should be noted that composition creativity was not evaluated in this study.

In contrast, the concept map drawn from a composition without the main idea does not contain a center with concrete meaning, as shown in Figure 6(b). The concept "myself" has the most relations, but it is meaningless and does not completely express the writer's position. This suggests that this kind of composition has a poor main idea, even though the composition is rich in ideas and development, so it will get a score below the grade of insufficient response according to the ATQE rubric and receive feedback such as "the position or main idea is very unclear" (excerpt from the ATQE rubric; all excerpts below are all from the ATQE rubrics). Figure 6(c) shows a concept map with weak main idea support because it has more local aggregation and fewer relations to the main concepts ("parents" and "myself") than Figure 6(a). Figure 6(c) is split into several parts, suggesting it will get a lower score for ideas and feedback such as "provides uneven support for the main idea." The other type of composition is one with vague ideas development, as seen in Figure 6(d). In this essay, the center is clear and rounded with rich concepts and relations; however, there are many similar concepts, and the relations are too dense, which indicates little variation between concepts and implies that the ideas are not fully developed. This type of essay will get feedback, such as "sentences and words may be simple and unvaried." The last one is the off-topic composition in Figure 6(e). Here, the key concept ("accompany") shared few connections with the rest of the concept map, and most of the other concepts make up a subset that is semantically far from the key concept. The result illustrates that the key concept is isolated from the main body of concepts, so it can be identified as an off-topic composition with feedback such as "provides no relevant ideas and content in response to the topic."

Providing meaningful feedback is of great importance for teachers and students. Systems with feedback can not only help to reduce the scoring load for teachers but can also encourage students to autonomously improve their writing ability. Therefore, the system AECC-I based on concept map can be used in the classroom or at home.

https://doi.org/10.1017/S1351324921000103 Published online by Cambridge University Press

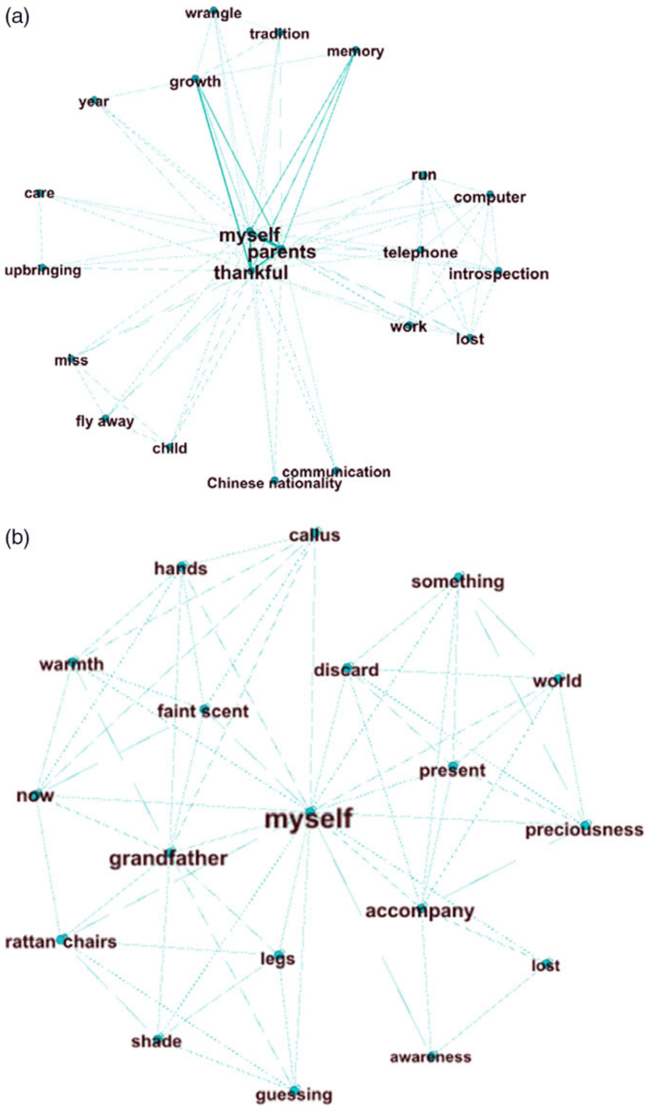

Figure 6. Concept maps made from compositions with idea qualities ranging from excellent to off-topic.

5. Discussion

5.1 General discussion

The findings in this study confirm a basic assumption that the quality of the ideas in essays can be captured by characteristics in concept maps via two aspects. Firstly, the proposed concept map features a framework that can predict the ideas score well, and secondly, more importantly, it also exhibits promising to cover the ideas scoring construct.

This study uses concept maps to supplement the higher-order trait of construct coverage of automated writing evaluations. Ideas of the internal cognition of students during writing are reflected through an intuitive conceptual network structure. Based on this concept map, this study can automatically score the ideas in a student's composition. Moreover, the concept map feedback is closely related to writing strategies, so it prioritizes writing skills rather than linguistic feedback (Stevens, Slavin and Farnish 1991).

https://doi.org/10.1017/S1351324921000103 Published online by Cambridge University Press

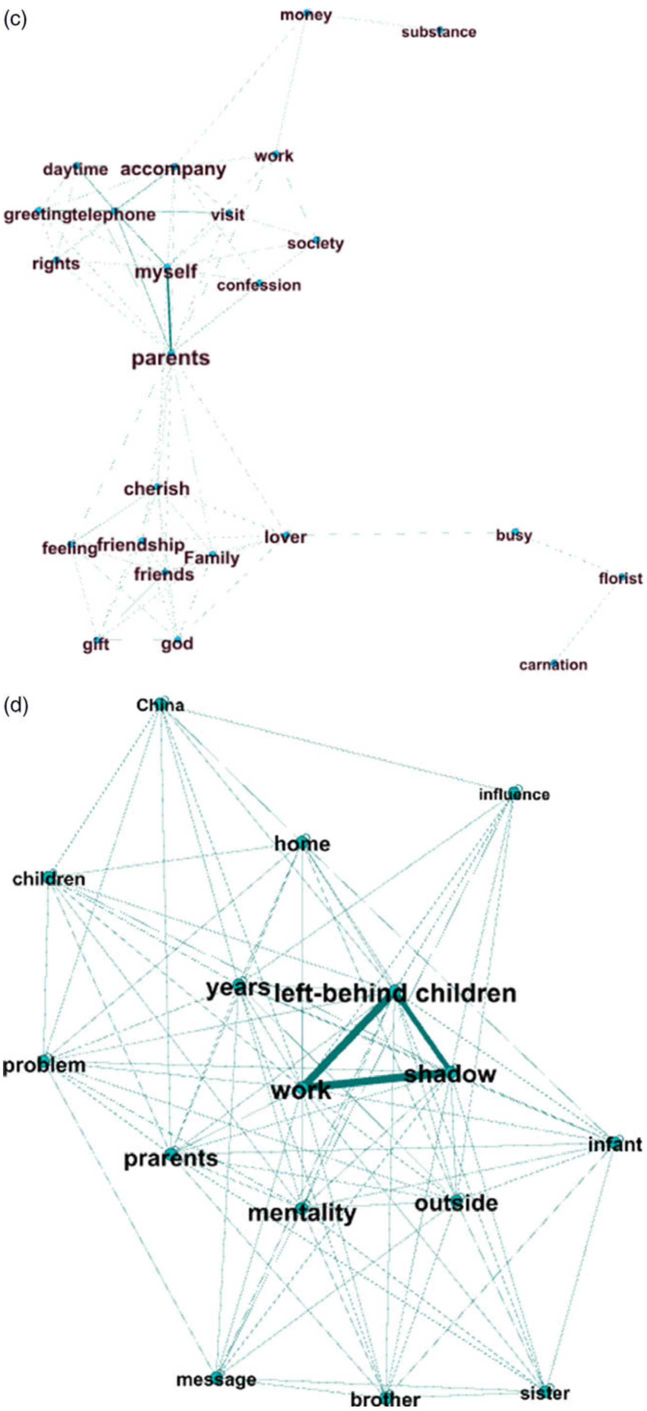

Figure 6. Continued.

https://doi.org/10.1017/S1351324921000103 Published online by Cambridge University Press

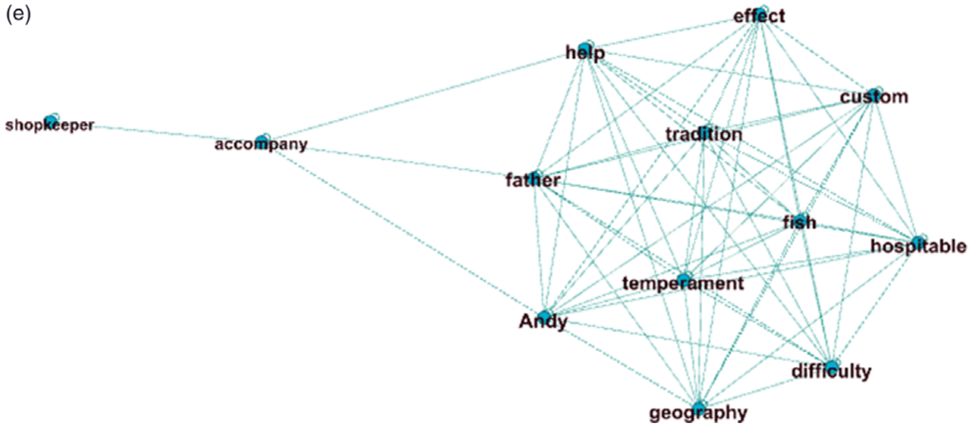

Figure 6. Continued.

Our research is largely inspired by Somasundaran *et al.* (2016) and Zupanc and Bosnic (2017). Somasundaran *et al.* (2016) only used two types of features derived from a graph: features based on degree and PR. These features are used to capture idea development and calculate the features without using embedded semantic representations. Our research uses four sets of features, and the first three feature sets are similar to those in Zupanc and Bosnic's research (2017). However, Zupanc and Bosnic (2017) used multiple features to predict the holistic scores rather than a specific dimension of ideas, and the nodes in the graph were sequentially overlapping on parts of the essay (a common method of measuring coherence); so, each node had only two neighbors in the graph. By contrast, in our concept map research, nodes are key concepts rather than word types (in Somasundaran *et al.* 2016) or sequence parts of an essay (in Zupanc and Bosnic 2017). In terms of semantic representation, we use word2vec embeddings instead of TF-IDF as in Zupanc and Bosnic's study (2017). Also, we proposed a group of similarity features, which was not included in previous studies, to access the ideas by estimating the similarity between an essay and high-scoring essays based on the graph. Furthermore, we expect the four groups of features to have the ability to capture the writing criteria of ideas and demonstrate the strengths and weaknesses of an essay (as illustrated in Figure 6). For example, it is more meaningful and easier to understand feedback that contains valuable insight rather than just pointing out semantic inconsistencies or grammar errors. These findings require further verification, but it is encouraging that the evaluation of ideas can be predicted using computational features based on a concept map created from a composition. To a certain extent, this study adopts a hypothesis-driven method and selects specific concept-based features that are assumed to aid in evaluating ideas. Our hypothesis method can be replaced by other theoretical hypotheses; however, judging from the research results so far, these hypotheses need to be further verified.

It is noted that this study hopes to distilled concepts that are as close as possible to the content of an essay so that each node in the graph can represent or be close to a concept rather than a word (Ke, Zeng and Luo 2016; Somasundaran *et al.* 2016). The process of concept recognition includes filtering stop words, coreference resolution, and merging synonyms. These recognitions are a key step to merge words that have the same (or very similar) semantics. Generally speaking, the more complex a concept is, the higher the level of abstraction, so we need to make a trade-off between complexity and generality to define a concept. This study takes word types that appear the most in the corpus and consider it as the standard word type to represent a synonym word group. As a result, we mainly used the word types close to the leaf nodes in the thesaurus tree as the concepts, so the degree of abstraction and complexity of the concepts are often not high. In the future, improving the degree of concept abstractions is a direction of our research, and we hope to investigate the relationship between highly abstract concepts and the quality of ideas.

https://doi.org/10.1017/S1351324921000103 Published online by Cambridge University Press

Moreover, this approach is it is not limited by genre or argument, which is very important for large Chinese writing assessment projects. The previous research on automatic scoring for Chinese compositions lacks a general and stable scoring feature system (Zhang and Ren 2004; ; Cao and Yang 2007; Xu *et al.* 2015; Liu, Qin and Liu 2016; Fu *et al.* 2018; Zhong and Zhang 2019), and the relationship between features and scoring criteria is vague or difficult to explain. This makes it difficult to support model validity and is not conducive to the application and promotion of automatic scoring systems in practice. The three datasets (prompts 1, 2, and 3) did not restrict genres (description, narration, and argumentation), with relatively broad and unclear topics, as students decide their composition's main idea and organizational flow. Therefore, the proposed model should be adaptable to a composition set with multiple writing focuses, rather than just evaluating the argument for a specific idea.

5.2 Relationship with AM-based automatic essay scoring

Although concept mapping is not a new technology, previous studies have shown that the local and global network structure of the text can indeed show interesting and different patterns between different types of individuals and groups (Qiao and Hu 2020). The focus of our work is to distinguish these differences and to rank them from high to low. Compared with AM research, this is a relatively robust rating. In the context of essay scoring, the profound significance of AM research lies in explaining "why" from an argumentative point of view. For example, which AM technology-based features base contribute to the persuasive strength of the composition, and how complex concepts constitute the argument structure. In fact, a well-written and persuasive essay often clearly states the author's point of view and supports their position by evoking relevant ideas and concepts. However, one of the main challenges of AEE is the inability to use specific prompt-specific models to evaluate new topics (Attali Bridgeman and Trapani 2010; Lee 2016), few researchers have made progress on general essay scoring (Mayfield and Black 2020). Unlike the AM method, the concept map-based method and BERT are better suited for new topics. Although the method in this study may be a simplified solution to access the quality of ideas compared with the argument mining-based method to access the quality of ideas, it may also reduce the complexity of identifying argument components and relationships, while still being able to investigate the "why" based on the argument mining. Considering the generalizability, timeliness, and fairness of language tests, the concept map-based method still has great potential. The starting points of the two research methods may be different, but the two methods are not incompatible. We are also very pleased to see that the latest research applies graphical representation to argument recognition (Dumani *et al.* 2021) and are looking forward to applying this method to further the automatic assessment of the quality of ideas research.

5.3 Comparison with automatic essay scoring based on BERT model

While neural networks have reduced the costs of feature engineering in automated essay scoring, the cost to fine-tune BERT models to achieve satisfactory performance has increased, especially with limited training data. To some extent, this may be a problem of value orientation; so the shift to deep neural models needs to take into consideration the price (Mayfield and Black 2020). In psychometrics, AEE researchers tend to give priority to simpler models, and the general concern is whether essay scoring is based on a set of strict and well-defined criteria (Attali 2013). For example, psychometricians select a set of reasonable variables for multiple regression (Attali and Burstein 2004). This model preserves the mapping between variables and identifiable dimensions, such as coherence or lexical complexity (Yannakoudakis and Briscoe 2012; Vajjala 2018). Neural networks definitely have great potential for automatic composition scoring, such as an intermediate representation of an essay (Fiacco, Cotos and Rose 2019; Nadeem *et al.* 2019), or in combination with handcrafted features (Peters, Ruder and Smith 2019). It can be inferred that

https://doi.org/10.1017/S1351324921000103 Published online by Cambridge University Press

neural models, such as BERT, have different scoring contributions for various aspects of an essay's quality. Therefore, in addition to the holistic score and ideas score, these models are also worth exploring for scoring other dimensions of an essay's quality.

5.4 Limitations

This study uses a scoring model based on the characteristics of concept maps to automatically score ideas, rather than the writing quality in Chinese compositions. We speculate that the characteristics derived from concept maps also relate to other writing aspects, but this method should be tested carefully before scoring other writing aspects.

This study is a starting point for evaluating automated scoring of ideas, as many potential research topics might improve upon it. For automatic assessing methods based on concept maps, one important consideration regards the method to construct the edges in concept maps. In this study, a co-occurrence relation was used to decide if there was an edge between two nodes. The assumption is that the co-occurrence relation between two concepts implies a grammatical or semantic relationship between them; future researchers can construct concept maps based on grammatical relations and semantic relations to capture the ideas relations and semantic distribution in compositions more effectively.

Lastly, AECC-I currently only assesses middle school written essays, so the stability and generalization of the system require further investigation, such as more reliability and validity studies because the model stability in different subgroups was not fully measured (such as fairness in relation to fair different grades, regions, and genders). Furthermore, although this study can predict composition idea scores, the features adopted may also relate to other dimensions of composition evaluation to some extent, such as composition structure and word choice. Lastly, the relationship between concept map features, other scoring dimensions, and total scores needs to be explored.

6. Conclusion

This study investigates multifaceted features based on concept maps derived from Chinese students' essays. When fusing the structure of a concept map and the semantic representation of word2vec to calculate the features, the handcrafted features method significantly better than BERT models and Coh-Metrix baselines. These results hopefully inspire future research into the potential benefits of developing the structural model of the knowledge graph in combination with trendy neural models.

The feature framework in this study has good structural validity, which can improve the transparency of automated scoring and can contribute to providing constructive feedback for students. The features extracted from these concept maps have a high-to-low correlation with idea scores, and factor analysis illustrates that the 12 features grouped into 4 factors (the main idea factor, a local support factor, a development factor, and a similarity factor) can explain the quality of the ideas more intuitively and concisely. Furthermore, the three features of global cohesion of ideas make the largest contribution to ideas scoring, which indicates stronger global cohesion; thus, higher idea scores. Lastly, a visual inspection of the concept maps can reveal distinctions between poor and excellent ideas in different compositions.

Acknowledgements. We thank Rui-Ji Fu and his team at Beijing research institute of Iflytek Co., Ltd. for the help on providing the toolkit of Chinese word segmentation and data-preprocess.

Financial support. We also thank the National Assessment Center for Education Quality of Chinese Ministry of Education for the financial support. The opinions expressed are those of the author and do not represent views of the Institute or the Department of Education.

https://doi.org/10.1017/S1351324921000103 Published online by Cambridge University Press

References

Aggarwal C.C. and Zhao P. (2013). Towards graphical models for text processing. *Knowledge & Information Systems* **36**, 1–21.

Amancio D.R., Oliveira Jr. **O.N. and Costa L.D.F.** (2012). Structure–semantics interplay in complex networks and its effects on the predictability of similarity in texts. *Physica A: Statistical Mechanics and Its Applications* **391**, 4406–4419.

Antiqueira L., Nunes M.G.V., Oliveira Jr. **O.N. and Costa L.D.** (2007). Strong correlations between text quality and complex networks features. *Physica A: Statistical Mechanics and Its Applications* **373**, 811–820.

Aryadoust V. and Liu S. (2015). Predicting EFL writing ability from levels of mental representation measured by Coh-Metrix: A structural equation modeling study. *Assessing Writing* **24**, 35–58.

Attali Y. (2007). On-the-*Fly Customization* of *Automated Essay Scoring*. ETS Research Report Series, 2, 2007–25.

Attali Y. (2011). Automated Subscores for TOEFL iBT® Independent Essays. ETS Research Report Series, 2011–39.

Attali Y. (2013). Validity and reliability of automated essay scoring. In **Shermis M.D. and Burstein J.C.** (eds), *Handbook of Automated Essay Evaluation: Current Applications and New Directions*, New York, NY: Routledge, pp. 181–198.

Attali Y. and Sinharay S. (2015). Automated *Trait Scores* for TOEFL® *Writing Tasks*. ETS Research Report Series, 2015(1).

Attali Y., Bridgeman B. and Trapani C. (2010). Performance of a Generic Approach in Automated Essay Scoring. *Journal of Technology, Learning, and Assessment* **10**. Available at https://www.learntechlib.org/p/106317/ (accessed 19 April 2021).

Attali Y. and Burstein J. (2004). *Automated Essay Scoring with E-Rater® v.2.0*. ETS Research Report Series, 2004(2), i–21.

Attali Y. and Burstein J. (2006). Automated essay scoring with e-rater® v.2. *Journal of Technology Learning & Assessment* **4**, i–21.

Bridgeman B. (2013). A simple answer to a simple question on changing answers. *Journal of Educational Measurement* **49**, 467–468.

Brin S. and Page L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems* **30**, 107–117.

Burstein J., Tetreault J. and Madnani N. (2013). The e-rater® automated essay scoring system. In **Shermis M.D. and Burstein J.C.** (eds), *Handbook of Automated Essay Evaluation: Current Applications and New Directions*, New York, NY: Routledge, pp. 55–67.

Cao Y. and Yang C. (2007). Chinese composition automatic scoring using latent semantic analysis. *Examination Research* **1**, 65–73. (In Chinese)

Chen N., Zhu J., Xia F. and Zhang B. (2015). Discriminative relational topic models. *IEEE Transactions on Pattern Analysis & Machine Intelligence* **37**, 973–986.

Chen Y. (2016). *Research on Chinese composition automatic scoring technology based on regression analysis. Master's degree*, Harbin Institute of Technology (In Chinese).

Cocarascu O. and Toni F. (2018). Combining deep learning and argumentative reasoning for the analysis of social media textual content using small data sets. *Computational Linguistics* **44**, 833–858.

Cohen J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin* **70**, 213–220.

Condon W. (2013). Large-scale assessment, locally developed measures, and automated scoring of essays: Fishing for red herrings? *Assessing Writing* **18**, 100–108.

Cui X. (2001). Comparison of Chinese and American composition evaluation criteria. *Chinese Teaching Communication* **22**, 28–29 (In Chinese).

Deane P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing* **18**, 7–24. doi: 10.1016/j.asw.2012.10.002.

Devlin J., Chang M., Lee K. and Toutanova K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the NAACL-HLT Conference, Minneapolis, MN, USA*, pp. 4171–4186.

Dumani L., Biertz M., Witry A., Ludwig A.K., Lenz M., Ollinger S., Bergmann R. and Schenkel R. (2021). The ReCAP Corpus: A Corpus of Complex Argument Graphs on German Education Politics. *IEEE 15th International Conference on Semantic Computing (ICSC), Laguna Hills, CA, USA, 2021*, pp. 248–255.

Eckle-Kohler J., Kluge R. and Gurevych I. (2015). On the role of discourse markers for discriminating claims and premises in argumentative discourse. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portuga*, pp. 2236–2242l.

Farra N., Somasundaran S. and Burstein J. (2015). Scoring persuasive essays using opinions and their targets. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications, Denver, CO, USA*, pp. 64–74.

Feng G. (1990). A brief discussion on the topic of the paper – and how to guide students to determine the central idea. *Journal of Kaifeng Institute of Education* **3**, 82–85 (In Chinese).

Fiacco J., Cotos E. and Rose C. (2019). Towards enabling feedback on rhetorical structure with neural sequence models. *In Proceedings of the 9th International Conference on Learning Analytics & Knowledge, Tempe, AZ, USA*, pp. 310–319.

Fleiss J.L. and Cohen J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement* **33**, 613–619.

Foltz P.W., Streeter L.E., Lochhaum K.E. and Landauer T.K. (2013). Implementation and applications of the intelligent essay assessor. *Journal of International Cooperation in Education* **15**, 159–168.

https://doi.org/10.1017/S1351324921000103 Published online by Cambridge University Press

Fu R., Wang D., Wang S., Hu G. and Liu T. (2018). Recognition of beautiful sentences for automatic composition grading. *Journal of Chinese Information* **32** (In Chinese).

Ghosh D., A. Khanam, Y. Han and S. Muresan (2016). Coarse-grained argumentation features for scoring persuasive essays. *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics,* Berlin, Germany, Volume 2: Short Papers, 549–554.

Ghosh D., Klebanov B.B. and Song Y. (2020). An exploratory study of argumentative writing by young students: A transformer-based approach. arXiv preprint arXiv:2006.09873.

Graesser A.C. and McNamara D.S. (2012). Automated analysis of essays and open-ended verbal responses. In *APA Handbook of Research Methods in Psychology*, Vol. **1**: Foundations, Planning, Measures, and Psychometrics. American Psychological Association, pp. 307–325. https://doi.org/10.1037/13619-017

Graesser A.C., McNamara D.S. and Kulikowich J.M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher* **40**, 223–234.

Graesser A.C., Mcnamara D.S., Louwerse M.M. and Cai Z. (2004). Coh-metrix: Analysis of text on cohesion and language. *Behavior Research Methods Instruments & Computers* **36**, 193.

Greene S. (1995). Making sense of my own ideas: The problems of authorship in a beginning writing classroom. *Written Communication* **12**, 186–218.

Higgins D., Burstein J. and Attali Y. (2006). Identifying off-topic student essays without topic-specific training data. *Natural Language Engineering* **12**, 145–159.

Janda H.K., Pawar A., Du S. and Mago V. (2019). Syntactic, semantic and sentiment analysis: The joint effect on automated essay evaluation. *IEEE Access* **7**, 108486–108503, doi: 10.1109/ACCESS.2019.2933354.

Ke X., Zeng Y., Ma Q. and Zhu L. (2014). Complex dynamics of text analysis. *Physica A: Statal Mechanics and Its Applications* **415**, 307–314.

Ke X., Zeng Y. and Luo H. (2016). Autoscoring essays based on complex networks. *Journal of Educational Measurement* **53**, 478–497.

Ke Z. (2019). *Automated Essay Scoring: Argument Persuasiveness,* Master's degree, The University of Texas at Dallas, USA.

Kim M. (2012a). Theoretically grounded guidelines for assessing learning progress: Cognitive changes in ill-structured complex problem-solving contexts. *Educational Technology Research and Development* **60**, 601–622.

Kim M. (2012b). Cross-validation study on methods and technologies to assess mental models in a complex problem solving situation. *Computers in Human Behavior* **28**, 703–717.

Kim M. (2013). Concept map engineering: Methods and tools based on the semantic relation approach. *Educational Technology Research & Development* **61**, 951–978.

Koszalka T.A. and Spector J.M. (2003). A review of two distance learning books [book review]. *Evaluation & Program Planning* **26**, 225–228.

Kwon N., Zhou L., Hovy E. and Shulman S. (2007). Identifying and classifying subjective claims. In Proceedings of the 8th Annual International Conference on Digital Government Research: Bridging Disciplines & Domains, Philadelphia, PA, USA, pp. 76–81.

Landauer T.K., Lochbaum K.E. and Dooley S. (2009). A new formative assessment technology for reading and writing. *Theory into Practice* **48**, 44–52.

Latifi S. and Gierl M. (2020). Automated scoring of junior and senior high essays using Coh-Metrix features: Implications for large-scale language testing. *Language Testing* **38**, 62–85.

Leckie G. and Baird J.A. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement* **48**, 399–418.

Lee Y. (2016). Investigating the feasibility of generic scoring models of E-rater for TOEFL iBT independent writing tasks. *English Language Teaching* **28**, 101–122.

Li B., Liu T., Qin B. and Li S. (2003). Chinese sentence similarity calculation based on semantic dependency. *Computer application research* **20**(12), 15–17. (In Chinese).

Li Y. (2006). *Chinese as a second language study to test the composition of automatic scoring*. Doctoral dissertation, Beijing, language and Culture University (In Chinese).

Li J., Zhou J., Luo X. and Yang Z. (2012). Chinese lexical networks: The structure, function and formation. *Physica A: Statistical Mechanics and Its Applications* **391**, 5254–5263.

Liu J., Xu Y. and Zhao L. (2019). Automated scoring based on two-stage learning. arXiv preprint. https://arxiv.org/pdf/1901.07744v1.pdf.

Liu M., Qin B. and Liu T. (2016). Automatic scoring of college entrance examination compositions based on literary features. *Intelligent Computers and Applications* **6**, 1–4. (In Chinese).

Louis A. and Higgins D. (2010). Off-topic essay detection using short prompt texts. In *P*roceedings of the NAACL HLT 2010 *Fifth Workshop* on *Innovative Use* of NLP for *Building Educational Applications,* Los Angeles, AK, USA, pp. 92–95.

Lumley T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing* **19**, 246–276.

Jin X. (1992). Sentence group discourse - unclear main idea. *Journalism and Writing* **6**, 28–30 (In Chinese).

Mayfield E. and Black A. (2020). Should You Fine-Tune BERT for Automated Essay Scoring? *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications,* Online, pp. 151–162. doi: 10.18653/v1/2020.bea-1.15.

Maharjan N. and Rus V. (2019). A concept map based assessment of free student answers in tutorial dialogues. In Isotani S., Millán E., Ogan A., Hastings P., McLaren B. and Luckin R. (eds), *Artificial Intelligence in Education*, AIED 2019. Lecture Notes in Computer Science, vol. 11625, Cham: Springer, pp. 244–257.

Martin B. and Eklund P. (2008) From concepts to concept lattice: A border algorithm for making covers explicit. In: Medina R. and Obiedkov S. (eds), *Formal Concept Analysis, ICFCA 2008.* Lecture Notes in Computer Science, vol. 4933. Berlin, Heidelberg: Springer.

Mayfield E. and Black A. (2020). Should you fine-tune bert for automated essay scoring? In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Online, pp. 151–162.

Meadows M. and Billington L. (2005). A review of the literature on marking reliability. *National Asessment Agency* (May), 89.

Mikolov T., Chen K., Corrado G.S. and Dean J. (2013). Efficient estimation of word representations in vector space. In Proceedings of Workshop at ICLR *2013, Scottsdale, AZ, USA*.

Mo M. (2018). Automated scoring for Chinese Composition. In The 13th Cross-Strait Conference on Psychological and Educational Assessment, Taizhong, China (In Chinese).

Mochales-Palau R. and Moens M. (2009). Argumentation mining: The detection, classification and structure of arguments in text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law, ICAIL 2009*, New York, NY, USA: ACM, pp. 98–107.

Nadeem F., Nguyen H., Liu Y. and Ostendorf M. (2019). Automated essay scoring with discourse-aware neural models. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, Florence, Italy, pp. 484–493.

Nafa F., Khan J.I., Othman S. and Babour A. (2016). Discovering bloom taxonomic relationships between knowledge units using semantic graph triangularity mining. In International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), Chengdu, China, pp. 224–233. doi: 10.1109/CyberC.2016.52.

National Writing Project (NWP). (2006). University of California, Berkeley, CA., America, Retrieved from http://www.writingproject.org/

Nguyen H. and Litman D. (2016). Context-aware argumentative relation mining. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany, Volume 1: Long Papers, pp. 1127–1137.

Nguyen H. and Litman D. (2018). Argument mining for improving the automated scoring of persuasive essays. In *Thirty-Second AAAI Conference on Artificial Intelligence*, LA, USA.

Ni Y., Bai K., Zhang X. and Liao C. (2013). Chinese text automation index construction and application - sentence minimum editing distance and structure similarity. In 19th *Information Management* and *Practice Conference*, *National Taichung University* of *Science* and *Technology*, Taibei, Taiwan (In Chinese).

Ni Y., Zhang X., Liao C., Guo B. and Bai K. (April 2014). Chinese text automatic analysis system is built based on real word pen, polysemy and stroke number. The 8th *International Symposium on information technology*, Taibei, Taiwan. (In Chinese).

NWREL. (2014). Traits Rubrics for Grade 3–12. Electronic pullet in board online. Available at http://educationnorthwest.org/traits/traits-rubrics

Ong N., Litman D. and Brusilovsky A. (2014). Ontology-based argument mining and automatic essay scoring. In *Proceedings of the First Workshop on Argumentation Mining*, Baltimore, MD, USA, pp. 24–28.

Opitz J. and Frank A. (2019). Dissecting content and context in argumentative relation analysis. In *Proceedings of the 6th Workshop on Argument Mining*, Florence, Italy, pp. 25–34, arXiv preprint arXiv:1906.03338.

Page E.B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan*, **47**, 238–243.

Pennington J., Socher R. and Manning C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, pp. 1532–1543.

Perelman L. (2012). Construct validity, length, score, and time in holistically graded writing assessments: The case against automated essay scoring. Bazerman C., Dean C., Early J., Lunsford K., Null S., Rogers P. and Stansell A. (eds), *International Advances in Writing Research: Cultures, Places, Measures,* The WAC Clearinghouse: Parlor Press, pp. 181–198.

Persing I. and Ng V. (2013). Modeling thesis clarity in student essays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, Volume 1: Long Papers, pp. 260–269.

Persing I. and Ng V. (2014). Modeling prompt adherence in student essays. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics,* Baltimore, MD, USA, Volume 1: Long Papers, pp. 1534–1543.

Persing I. and Ng V. (2015). Modeling argument strength in student essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, Beijing, China, Volume 1: Long Papers, pp. 543–552.

Persing I. and Ng V. (2016). Modeling stance in student essays. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics,* Berlin, Germany, Volume 1: Long Papers, pp. 2174–2184.

https://doi.org/10.1017/S1351324921000103 Published online by Cambridge University Press

Persing I. and Ng V. (2020). Unsupervised argumentation mining in student essays. In *Proceedings of The 12th Language Resources and Evaluation Conference,* Marseille, France, pp. 6795–6803.

Peters M.E., Ruder S. and Smith N.A. (2019). To tune or not to tune? adapting pretrained representations to diverse tasks. In *Proceedings of the Workshop on Representation Learning for NLP* (RepL4NLP-2019), Florence, Italy, pp. 7–14.

Pirnay-Dummer P. and Ifenthaler D. (2010). Automated knowledge visualization and assessment. In: Ifenthaler D., Pirnay-Dummer P. and Seel N. (eds), Boston, MA, USA: Springer.

Powers D.E., Burstein J., Chodorow M., Fowles M.E. and Kukich K. (2002). Stumping e-rater: Challenging the validity of automated essay scoring. *Computers in Human Behavior* **18**, 103–134.

Qiao C. and Hu X. (2020). A neural knowledge graph evaluator: Combining structural and semantic evidence of knowledge graphs for predicting supportive knowledge in scientific QA. *Information Processing & Management* **57**, 102309.

Quinlan T., Higgins D. and Wolff S. (2009). Evaluating the construct-coverage of the e-rater®; Scoring engine. *ETS Research Report*, **2009**, i–35.

Rahimi Z., Litman D., Correnti R., Wang E. and Matsumura L. (2017). Assessing students' use of evidence and organization in response-to-text writing: Using natural language processing for rubric-based automated scoring. *International Journal of Artificial Intelligence in Education* **27**, 694–728.

Ramineni C. and Williamson D.M. (2013). Automated essay scoring: Psychometric guidelines and practices. *Assessing Writing* **18**, 25–39.

Rei M. (2017). Detecting off-topic responses to visual prompts. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, Copenhagen, Denmark, pp. 188–197.

Rodriguez P., Jafari A. and Ormerod C. (2019). Language models and automated essay scoring. arXiv preprint arXiv:1909.09482.

Ruegg R. and Sugiyama Y. (2013). Organization of ideas in writing: What are raters sensitive to? *Language Testing in Asia* **3**, 8.

Rumelhart D.E. 1986. Learning internal representations by error propagation. In Rumelhart D.E., Mccelland J.L. and PDP Research Group (eds), *Parallel Distributed Processing*, 1.

Saint-Dizier P. (2017). Knowledge-driven argument mining based on the qualia structure. *Argument & Computation* **8**, 193–210.

Shermis M.D. (2002). Trait rating for automated essay grading. *Educational & Psychological Measurement* **62**, 5–18.

Shermis M.D., Burstein J. and Bursky S. (2013). Introduction to automated essay evaluation. In Shermis M.D. and Burstein J. (eds), *Handbook of Automated Essay Evaluation: Current Applications and New Directions.* New York: Routledge, 2013, p1–15.

Shermis M.D., Burstein J., Higgins D. and Zechner K. (2010). Automated essay scoring: Writing assessment and instruction. *International Encyclopedia of Education*, 20–26.

Shin J. and Gierl M. (2020). More efficient processes for creating automated essay scoring frameworks: A demonstration of two algorithms. *Language Testing.* https://doi.org/10.1177/0265532220937830.

Schlomske N. and Pirnay-Dummer P. (2008). Model based assessment of learning dependent change during a two semester class. In Kinshuk, S.D. and Spector M. (eds.), *Proceedings of* IADIS *International Conference Cognition and Exploratory Learning in Digital Age 2008,* Freiburg, Germany: IADIS, pp. 478–480.

Spandel V. and Stiggins R.J. (1990). Creating writers: Linking assessment and writing instruction. In *College Composition and Communication*, New York: Longman, pp. 478–480.

Stab C. and Gurevych I. (2014a). Annotating argument components and relations in persuasive essays. In *Proceedings of the 25th International Conference on Computational Linguistics* (COLING 2014), Dublin, Ireland, pp. 46–56.

Stab C. and Gurevych I. (2014b). Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP),* Doha, Qatar, pp. 46–56

Stab C. and Gurevych I. (2017). Parsing argumentation structures in persuasive essays. *Computational Linguistics* **43**, 619–659.

Stevens R.J., Slavin R.E. and Farnish A.M. (1991). The effects of cooperative learning and direct instruction in reading comprehension strategies on main idea identification. *Journal of Educational Psychology* **83**, 8–16.

Somasundaran S., Riordan B., Gyawali B. and Yoon S. (2016). Evaluating Argumentative and Narrative Essays using Graphs. In *The 26th International Conference on Computational Linguistics,* Osaka, Japan, Technical Papers, pp. 1568–1578.

Song Y., Heilman M., Klebanov B. and Deane P. (2014). Applying argumentation schemes for essay scoring. In P*roceedings of the First Workshop on Argumentation Mining* Baltimore, Association for Computational Linguistics. Baltimore, MD, USA, pp. 69–78.

Vajjala S. (2018). Automated assessment of non-native learner essays: Investigating the role of linguistic features. *International Journal of Artificial Intelligence in Education* **28**, 79–105.

Villalon J. and Calvo R.A. (2011). Concept maps as cognitive visualizations of writing assignments. *Journal of Educational Technology & Society* **14**, 16–27.

https://doi.org/10.1017/S1351324921000103 Published online by Cambridge University Press

Wachsmuth H., Khatib K. and Stein B. (2016). Using argument mining to assess the argumentation quality of essays. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics,* Osaka, Japan, Technical Papers, pp. 1680–1691.

Wu Z. and Palmer M., (1994). Verb semantics and lexical selection. In *Acl Proceedings of Annual Meeting on Association for Computational Linguistics,* NM, USA, pp.133–138.

Xia L., Wen Q. and Pan K. (2017). Unsupervised off-topic essay detection based on target and reference prompts. In *13th International Conference on Computational Intelligence and Security* (CIS), Hong Kong, China, 2017. pp. 465–468. doi: 10.1109/CIS.2017.00108.

Xu C., Chen D., Wu Q. and Xie Z. (2015). Chinese as a second language composition automatic grading research. The International Chinese Teaching Research 1, 83–89 (In Chinese).

Yannakoudakis H. and Briscoe T. (2012). Modeling coherence in ESOL learner texts. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP,* Montréal, Canada, pp. 33–43.

Zedelius C.M., Mills C. and Schooler J.W. (2018). Beyond subjective judgments: Predicting evaluations of creative writing from computational linguistic features. *Behavior Research Methods* **51**, 879–894. doi: 10.3758/s13428-018-1137-1.

Zhang H. and Litman D. (2017). Word embedding for response-to-text assessment of evidence. In *Proceedings of ACL 2017*, Student Research Workshop, Vancouver, Canada, pp. 75–81.

Zhang H. and Litman D. (2018). Co-attention based neural network for source-dependent essay scoring. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, New Orleans, LA, USA, pp. 399–409.

Zhang H., Magooda A., Litman D., Correnti R., Wang E., Matsmura L.C., Howe E. and Quintana R. (2019). eRevise: Using natural language processing to provide formative feedback on text evidence usage in student writing. In *Proceedings of the AAAI Conference on Artificial Intelligence,* Hawaii, USA, Vol. 33, pp. 9619–9625.

Zhang J. and Ren J. (2004). Experimental research report of e-grader for Chinese language test. *Chinese Examinations* **10**, 27–32 (In Chinese).

Zhang X., Ni Y., Liao C., Kuo B. and Bai K. (2014). Content word strokes, word polysemy and stroke number index are used to build an automatic Chinese text analysis system. In *8th International Symposium on Information Technology* (In Chinese).

Zhong Q. and Zhang J. (2019). Embedded in language depth perception of Chinese essay scoring algorithm. *Computer Engineering and Application,* 2019 (In Chinese).

Zhou S., Hu G., Zhang Z. and Guan J. (2008). An empirical study of Chinese language networks. *Physica A: Statistical Mechanics and Its Applications* **387**, 3039–3047.

Zhou J., Luo Y. and Chen B. (2018). Research on thematic expression of syntactic subject. Application of Language 001, 61–70 (In Chinese).

Zouaq A., Gasevic D. and Hatala M. (2011). Towards open ontology learning and filtering. *Information Systems* **36**, 1064–1081.

Zouaq A. and Nkambou R. (2008). Building domain ontologies from text for educational purposes. *IEEE Transactions on Learning Technologies* **1**, 49–62.

Zupanc K. and Bosnic Z. (2017). Automated essay evaluation with semantic analysis. *Knowledge-Based Systems* **120**, 118–132.

Appendix A. An English translation of prompt 1.

Original:

请以《陪伴是最好的礼物》为题，写一篇文章。

要求：写作的文体不限，有鲜明的中心思想，充实的内容和流畅的语言表达，并且不少于400字。（6分）

Translation:

Please write a composition entitled “Companionship is the Best Gift.”

Requirement: The writing style is not limited, with a clear central idea, substantial content, and fluent language expression, and no less than 400 words. (6 marks)

https://doi.org/10.1017/S1351324921000103 Published online by Cambridge University Press

Appendix B. The baseline of the Chinese Coh-Metrix feature set (Table E1).

Words features	Explanation
Number of characters	Total number of text characters in the text
Number of words	Total number of words in the text
Percentage of one-character words	The ratio of the number of words consisting of one character to the total number of words
Percentage of two-character words	The ratio of the number of words consisting of two characters to the total number of words
Percentage of three-character words	The ratio of the number of words consisting of three characters to the total number of words
Percentage of more than four characters words	The ratio of the number of words consisting of four or more characters to the total number of words
Average number of strokes	The average stroke number of a single character in a full text.(Chinese characters are made up of simple strokes. The strokes of a pen or pencil are the movements that you make with it when you are writing Chinese characters (e.g., the character ‘天’ (sky) has four strokes))
Low stroke ratio	The number of characters with 1–10 strokes divided by the number of full-text characters
Medium stroke ratio	The number of characters with 11–20 strokes divided by the number of full-text characters
High stroke ratio	The number of characters with more than 21 strokes divided by the number of full-text characters
Average number of words in a sentence	The total number of words in the text divided by the number of sentences
Number of content words	The number of all content words in the full text
Number of nouns	The number of all nouns in the full text
Number of verbs	The number of all verbs in the full text
Number of adjectives	The number of all adjectives in the full text
Number of numerals	The number of all numerals in the full text
Number of quantifiers	The number of all quantifiers in the full text
Number of pronouns	The number of all pronouns in the full text
Number of interjections	The number of all interjections in the full text
Number of function words	The number of all function words in the full text
Number of prepositions	The number of all prepositions in the full text
Number of conjunctions	The number of all conjunctions in the full text
Number of particles	The number of all particles in the full text
Number of modal particles	The number of all modal particles in the full text
Number of interrogative pronouns	The number of all modal particles in the full text

https://doi.org/10.1017/S1351324921000103 Published online by Cambridge University Press

Frequency of pronouns	The logarithm frequency of pronouns in the full text
Frequency of content words	The logarithm frequency of content words in the full text
Frequency of function words	The logarithm frequency of function words in the full text
Frequency of modal particles	The logarithm frequency of modal particles in the full text
Frequency of interrogative pronouns	The logarithm frequency of interrogative pronouns in the full text
Frequency of verbs	The logarithm frequency of verbs in the full text
Frequency of adjectives	The logarithm frequency of adjectives in the full text
Frequency of prepositions	The logarithm frequency of prepositions in the full text
Frequency of conjunctions	The logarithm frequency of conjunctions in the full text
Connectives features	**Explanation**
Parataxis	The frequency of parataxis conjunctions
Progressive	The frequency of progressive conjunctions
Selective	The frequency of selective conjunctions
Succession	The frequency of succession conjunctions
Transition	The frequency of transition conjunctions
Hypothetical	The frequency of hypothetical conjunctions
Causal	The frequency of causal conjunctions
Condition	The frequency of condition conjunctions
Purpose	The frequency of purpose conjunctions
Total connectives	The frequency of total connectives conjunctions
Lexical diversity features	**Explanation**
TTR of content words	The ratio of the number of content word types to the number of tokens appearing in the full text
TTR of total words	The ratio of the number of word types to the number of tokens appearing in the full text
Average word length	The average word length in the full text
Sentences complexity	**Explanation**
Number of sentences	Total number of sentences in the full text
Average sentence length	The average number of characters in a sentence in the full text
Short sentence length	The number of characters in the shortest sentence in the full text
Long sentence length	The number of characters in the longest sentence in the full text
Average number of sentences per hundred words	The average number of sentences per hundreds of words in a full text

Sentence structure	
The minimum edit distance between adjacent sentences	The average of the minimum edit distance between the words of adjacent sentences in the text (modified the algorithm of Levenshtein (1965) to adapt to Chinese)
The minimum edit distance between words in the text	The average of the minimum edit distance between all words in the text (idem)
The minimum edit distance between adjacent parts of speech	The average of the minimum edit distance between parts of speech in adjacent sentences in the text (idem)
The minimum edit distance between the parts of speech in the text	The average of the minimum edit distance between all parts of speech in the text (idem)
Structural similarity of adjacent sentences	The similarity of adjacent sentences based on the dependency structure (Li *et al*. 2003)
Similarity of sentence structure in the whole text	The similarity of all sentences based on the dependency structure (Li *et al*. 2003)
Connection between sentences	**Explanation**
Content word overlap of adjacent sentences	The proportion of adjacent sentence pairs sharing one or more of the same content words to the total number of sentences
Content word overlap of all sentences	The proportion of all sentence pairs that share the same content word to the total number of sentences
Noun overlap of adjacent sentences	The proportion of adjacent sentence pairs sharing one or more of the same nouns to the total number of sentences
Noun overlap of all sentences	The proportion of all sentence pairs that share the same noun to the total number of sentences
Verb overlap of adjacent sentences	The proportion of adjacent sentence pairs sharing one or more of the same verbs to the total number of sentences
Verb overlap of all sentences	The proportion of all sentence pairs that share the same verb to the total number of sentences
LSA features	**Explanation**
LSA sentence adjacent	The average of the LSA cosine between adjacent sentences
LSA sentence all	The average of the LSA cosine between all sentences in the text
LSA verb overlap	The average of the verb LSA cosine between adjacent sentences
LSA givenness versus newness	The indicator of the information (as opposed to new information) exists in each sentence in a text, as compared with the content of prior text information.

Cite this article: Yang L-P, Xin T, Luo F, Zhang S and Tian X-T (2022). Automated evaluation of the quality of ideas in compositions based on concept maps. *Natural Language Engineering* **28**, 449–486. https://doi.org/10.1017/S1351324921000103