Expressing Metaphorically, Writing Creatively: Metaphor Identification for Creativity Assessment in Writing

Dongyu Zhang
School of Software
Dalian University of Technology
Dalian, China
zhangdongyu@dlut.edu.cn

Minghao Zhang
School of Software
Dalian University of Technology
Dalian, China
zhang.minghao@outlook.com

Ciyuan Peng
School of Engineering, IT and Physical Sciences
Federation University Australia
Ballarat, Australia
sayeon1995@gmail.com

Feng Xia*
School of Engineering, IT and Physical Sciences
Federation University Australia
Ballarat, Australia
f.xia@ieee.org

ABSTRACT

Metaphor, which can implicitly express profound meanings and emotions, is a unique writing technique frequently used in human language. In writing, meaningful metaphorical expressions can enhance the literariness and creativity of texts. Therefore, the usage of metaphor is a significant impact factor when assessing the creativity and literariness of writing. However, little to no automatic writing assessment system considers metaphorical expressions when giving the score of creativity. For improving the accuracy of automatic writing assessment, this paper proposes a novel creativity assessment model that imports a token-level metaphor identification method to extract metaphors as the indicators for creativity scoring. The experimental results show that our model can accurately assess the creativity of different texts with precise metaphor identification. To the best of our knowledge, we are the first to apply automatic metaphor identification to assess writing creativity. Moreover, identifying features (e.g., metaphors) that influence writing creativity using computational approaches can offer fair and reliable assessment methods for educational settings.

CCS CONCEPTS

• **Applied computing** → *Computer-assisted instruction*; • **Computing methodologies** → *Information extraction.*

KEYWORDS

Writing creativity assessment, Metaphor identification, Writing analytics, Textual data mining, Metaphorical feature analytics.

ACM Reference Format:
Dongyu Zhang, Minghao Zhang, Ciyuan Peng, and Feng Xia. 2022. Expressing Metaphorically, Writing Creatively: Metaphor Identification for Creativity Assessment in Writing. In *Companion Proceedings of the Web Conference 2022 (WWW '22 Companion), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3487553.3524935

1 INTRODUCTION

The English language has a global spread, and it is one of the most important and widely spoken international languages today. There is an increasingly large population of English learners who use English for professional purposes or for daily communication throughout the world. Language assessment, therefore, plays an essential role, with a massive number of English learners in the context of education [1, 2]. However, language assessment, particularly writing quality, requires intensive labor from assessors, who need adequate knowledge and extensive training in developing judgments and scoring rubrics, criteria, etc. [17]. Indeed, human raters tend to be subjective, and it is very challenging for them to be highly consistent with each other. Therefore, if we can standardize the score of writing quality, it will greatly help the language education work.

With the development of Natural Language Processing (NLP) techniques, a surge of research on Automated Essay Scoring (AES) has taken place to address high-volume workloads, costs, reliability across raters, and the need for timely feedback in manual writing assessments [12]. AES focuses on automatically predicting and scoring the quality of essays by using NLP and machine learning techniques [14]. However, machine-based assessment has been criticized for focusing on the surface linguistic mechanisms of writing rather than creativity (aesthetic and imaginative features of writing), although creativity is likely to contribute to high-quality writing [3].

Scholars link the use of figurative language, particularly metaphor, with the creativity of writing [16]. The nature of metaphor involves the cognitive process of conceptualizing and constructing novel terms by comparing the semantic similarities of two concepts [19, 25]. Specifically, one familiar, concrete concept is metaphorically used to view another more abstract, novel, and complex concept [20, 27]. This suggests that the appropriate employment of

[*] Corresponding author.

WWW '22 Companion, April 25–29, 2022, Virtual Event, Lyon, France.
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9130-6/22/04…$15.00
https://doi.org/10.1145/3487553.3524935

metaphor could contribute to creativity in the way that it creates vivid, novel, and imaginative thoughts and expressions [26].

Therefore, to assess the creativity of essays in writing assessment systems, metaphor identification is essential. In this paper, we propose a creativity assessment model based on automatic metaphor identification. When our creativity assessment model scores the creativity of each text, the metaphorical expression in the text is an important indicator. We first detect metaphors in texts by using a novel token-level metaphor identification method. The existing metaphor identification methods are phrase-level [4, 23] or token-level metaphor recognition method [24]. Nevertheless, the accuracy of token-level metaphor recognition methods is low. Also, existing methods rarely take advantage of the deep semantic information in texts and cannot establish a valid mapping between literal and metaphorical texts. Thus, this paper proposes a novel token-level metaphor identification method based on pre-training of deep bidirectional transformers (BERT) [6], Bi-Gated Recurrent Unit (Bi-GRU) [10] and Conditional Random Field (CRF) [13]. After detecting and extracting metaphors, we apply the extracted metaphor features to the creativity assessment model to detect creative writing. Our contributions are as follows:

- We propose a new metaphor identification-based writing creativity assessment model, which considers metaphorical expression as an important indicator of writing creativity scoring. Our model outperforms the state-of-the-art method. To the best of our knowledge, we are the first to use automatic metaphor identification to assess writing creativity.
- To detect metaphors accurately, this paper proposes a novel token-level metaphor identification method based on BERT, Bi-GRU, and CRF. Our algorithm has better performance on the F1-score than other token-level metaphor identification algorithms.
- We present NLP techniques addressing high-volume workloads and cross-rater reliability in creativity assessments, and we attempt fair and reliable assessments.
- We contribute to a novel dataset, which is being released publicly, with manually added annotation for each essay to measure creativity.

The rest of the paper is organized as follows. Section 2 describes our metaphor identification method. Section 3 presents the metaphor identification experiments. Section 4 describes our writing creativity assessment model. Section 5 gives the experimental results of creativity assessment. We conclude the paper in Section 6.

2 METAPHOR IDENTIFICATION METHOD

This paper uses a combination of BERT, Bi-GRU, and CRF to identify token-level metaphors in sentences. BERT is a pre-trained language model, which can effectively use context information to extract the relationship between ontology and metaphor because the transformer can notice multiple key points [9]. In this paper, we first use the word embedding function of BERT to convert sentences into a matrix. Then, the sentence matrix is inputted into the Bi-GRU model, which can efficiently learn and process past and future information in the sequence. Here, GRU is a special recurrent neural network (RNN) that learns long-term dependencies to overcome the issue of vanishing gradients in RNN[11]. By training Bi-GRU, the relationship between ontology and metaphor extracted by BERT can be learned, and a mapping between literal and metaphor can be established. Finally, the properties of each frame in the sequence are predicted by CRF, that is, predicting the probability of each word being a metaphorical word.

Figure 1 shows the main structure of the metaphor identification method proposed in this paper. BERT converts words into vectors and passes them to Bi-GRU. w_i indicates the i-th word in the text, l_i represents the i-th word and its left context information, r_i represents the i-th word and its right context information. c_i represents the concatenating two vectors of w_i in its context. BERT can extract text features and pass them to Bi-GRU. After training, Bi-GRU can learn the correlation and difference between ontology and metaphor. Finally, through the calculation of CRF, the model outputs the probability that each word in the sentence is metaphorical.

For example, "you have shipwrecked my career" is the input to the model. First, the words in the sentence are encoded by the embedded layer. The encoded representation of the word is then learned by the Transformer encoder. In this process, the text features and the metaphor information contained in them are extracted. The word code of the BERT output is passed to Bi-GRU as an input to it. The trained Bi-GRU can map the ontology to the metaphor. Finally, through the calculation of CRF, it can be determined that "shipwrecked" is a metaphorical word, and the others are literal. In this sentence, "shipwreck" means "damage", it is a typical metaphor.

3 METAPHOR IDENTIFICATION EXPERIMENTS

3.1 Datasets and Baseline

We use two datasets in the metaphor identification experiment. The first one is VU Amsterdam Metaphor Corpus (VUAMC) [1], which has been widely used in the study of metaphor computing. VUAMC is currently the largest hand-labeled metaphorical corpus. It includes news genres, academic texts, novels, and conversations, with a scale of 200,000 English words. The second one is Mohammad dataset[2], which is widely used for metaphor identification research [5, 18]. It contains 1,230 literal and 409 metaphor sentences. Each sentence contains a target word and its label, which is annotated by 10 annotators.

We select the results of Pramanick et al.[21] as one of our baselines. This work used Long-Short Term Memory (LSTM) and CRF to recognize token-level metaphors. They not only used the token features but also considered lexical features, for example, the lemma of the token, part-of-speech, and so on. We also compare with the results of Mao et al.[15]. They proposed an unsupervised learning method that identifies and interprets metaphors at token-level without any preprocessing, outperforming in the metaphor identification task. For both baseline models, we set parameters according to the original paper.

[1] http://www.vismet.org/metcor/documentation/home.html
[2] http://saifmohammad.com/

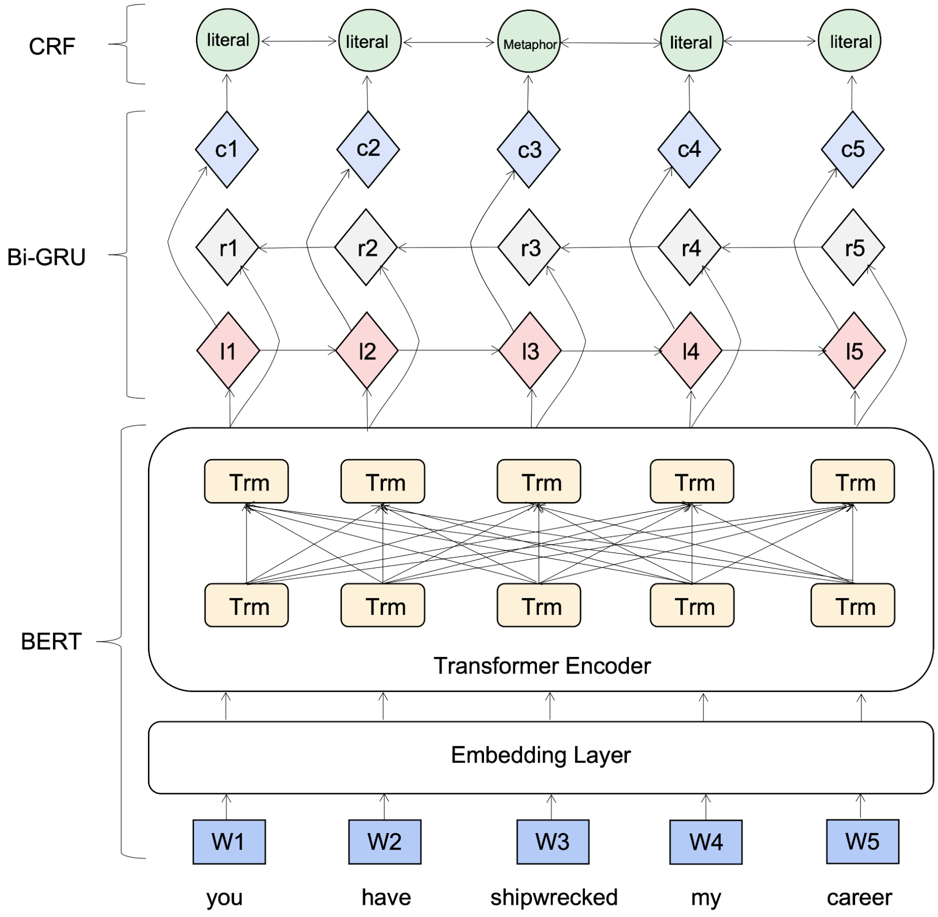

Figure 1: Main architecture of our metaphor identification method.

Table 1: Metaphor identificaiton over two test datasets.

Methods	VUAMC			Mohammad Dataset		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Pramanick et al.[21]	0.7036	0.5755	0.6327	0.6585	0.6126	0.6222
Mao et al.[15]	0.6837	0.7202	0.7045	0.6294	0.6397	0.6345
BERT	0.7433	0.7718	0.7584	0.7138	0.5875	0.6435
Bi-GRU	0.6529	0.5271	0.5832	0.7023	0.5288	0.6019
CRF	0.6082	0.7001	0.6514	0.6594	0.5516	0.6007
BERT+Bi-GRU	0.6631	0.7812	0.7225	0.8472	0.5571	0.6721
BERT+CRF	0.7534	**0.8431**	0.7912	0.7512	0.6024	0.6769
Bi-GRU+CRF	0.6825	0.7134	0.6973	0.5663	**0.6786**	0.6173
BERT+Bi-GRU+CRF	**0.8891**	0.8069	**0.8403**	**0.8872**	0.6321	**0.6929**

Table 2: Metaphor features used in our experiments.

Feature Name	Student Writing	News	Journal Paper	Webis-CPC-11
Title metaphor	6.3517#	-	-	-
Total words metaphor	0.0622	-0.6437	0.2332	0.5217
Token metaphor rate	0.0801	-0.3215	0.0656	0.1328
Metaphor pre sentence	0.0853	0.7364	0.0098	-0.3614
Metaphor pre paragraph	0.1206	-	-	-

3.2 Experimental settings

We consider all tokens regardless of the POS tags. We ignore punctuation marks such as commas (,), exclamation points (!), periods (.). After removing all punctuation marks, each token is marked as negative or positive, representing literal and metaphor, respectively. Since the length of each sentence is not necessarily the same, we pad them with zero vectors, unifying the sentence length to 50. We divided the data into a training set, development set, and test set with a ratio of 8:1:1.

Our model uses the pre-trained BERT model (BERT-Base, Uncased)[3] provided by Google research to extract text features. It contains 12 layers, 768 hidden neurons, 12 heads attention mechanism, and 110M parameters. We use a batch size of 128 during training and a learning rate of 0.001. We use Adam as our optimizer with a dropout rate of 0.2. Our model uses two single GRU layers for forwarding and backward propagation, respectively. The size of each layer is 200.

3.3 Experimental Results and Discussions

Table 1 shows the metaphor identification results of our model and baselines on two datasets. Our model obtains F1-scores of 0.8403 on VUAMC and 0.6929 on Mohammad et al.'s data set. As Table 1 shows, our method performs significantly better on two different data sets than the baselines. The performance of our model on VUAMC is significantly better than that on Mohammad et al.'s data set. The reason is that the size of train set of VUAMC is much larger than train set of Mohammad et al.'s data set, Therefore the model can learn more useful information on the training data with larger data sizes and improve the effect and accuracy of the classification. Through experimental verification, the model proposed in this paper is indeed effective and achieved better results than the baseline on the two data sets. At the same time, the generalization ability of the model is proved.

The results on the two data sets demonstrate the validity of our model. The pre-trained BERT model can effectively use context information to extract the relationship between ontology and metaphor. Bi-GRU can learn the relationship between ontology and metaphor extracted by BERT, and establish a mapping between literal and metaphor. Through the calculation of CRF, we can get the probability that if the target word being a metaphor. The new model can effectively identify the token-level metaphor in the sentence sequence.

4 CREATIVITY ASSESSMENT MODEL

4.1 Datasets

4.1.1 Student Writing Data. We first collect student writing data, which is from the British Academic Written English Corpus (BAWE) [4], comprising 2,593 pieces of proficient, assessed writing from 35 diverse disciplines. We add manual creativity scores from four popular subscales measuring creativity: fluency, originality, elaboration, and resistance to premature closure [22]. We apply 4 points to score essays (1 = least creative; 4 = most creative). We obtain fluency scores by counting supporting ideas. Originality scores are obtained from the probability of the thesis and the evidence. According to Pareto's law [7], the essential constituents of any group of things comprise only a small part, about 20%. The remaining 80%, although the majority, are secondary constituents. Therefore, if the perspective of a thesis is unique, the probability of it occurring elsewhere is less than 20%, and we assign 4 points; if the probability is more than 20%, we assign 1-3 points according to its rarity. Elaboration scores come from the number of added ideas. If there are more than three extra ideas, we assign 4 points. We score resistance to premature closure by examining the degree of psychological openness. The final score is the average of the four measures.

Four native English teachers with more than three-year experience in English language teaching score the writing. They are in two groups, with two in each group. We adopt the average score of the two scorers as the final score for the essay. However, when the two scorers give scores that differed by 2 points or more, we ask the other group to score the essay, and we used the average score of the two groups as a result. We use the kappa score statistic, κ, to measure agreement on the reliability of the scoring scheme. We score 100 essays from the dataset for creativity and agreement. We find that $\kappa = 0.89$, so the assessment was reliable.

4.1.2 Paraphrases, News and Academic Paper Data. Measuring the creativity of an article effectively is crucial. Creativity is a complex, multi-faceted concept. Thus, we also use online news, research journals, and paraphrases to test our model. We first use Webis Crowd Disphrase Corpus 2011 (Webis-CPC-11)[5] that consists of 4,067 accepted paraphrases, 3,792 rejected non-paraphrases and original text. Then, we collect various news articles (1,221 pieces) from news websites Onion [6] (a website proving satiric news), on which the news articles are more creative than other news articles. We regard the articles from the Onion as creative samples and the rest as non-creative samples. We exclude articles with fewer than five sentences, and this left 534 creative articles and 403 non-creative articles. We also collect an academic paper dataset to validate our creativity detection method. The more influential journals, the more influential the journal, the more important novelty becomes as a requirement. We select papers from ten influential journals which explicitly require high novelty for publication as creative samples (404 papers); and from nine low-influence journals (impact factor < 1.0) which do not explicitly require high novelty for publication as non-creative samples (496 papers) in mathematics, physics, computer science, bio-science, and chemistry.

4.2 Method

4.2.1 Metaphor Feature Extraction. We design a group of metaphor-wise features and test their performance in creativity detection. We use the metaphor detection method described above to find all possible metaphorical words in every sentence. Then, we compute the number of metaphorical words, the proportion of metaphorical words to target words, and the number of metaphorical words per sentence for each document in the datasets. We also calculate the number of metaphorical words per paragraph and metaphor in the title for documents in the student writing dataset. All the metaphor-wise features are in Table 2, wherein each row denotes that this

[3] https://github.com/google-research/bert#pre-trained-models

[4] http://ota.ox.ac.uk/desc/2539

[5] https://webis.de/data/webis-cpc-11.html

[6] https://www.theonion.com/

Table 3: Performance of creativity classification models with different feature vectors.

Feature Vector	Standard	Student Writing	News	Journal Paper	Webis-CPC-11
Document feature vector	Accuracy	0.6517	0.9101	0.7122	0.7963
	F1-score	0.4623	0.9218	0.6534	0.7715
Metaphor feature vector	Accuracy	0.6224	0.8811	0.6722	0.7280
	F1-score	0.3623	0.8967	0.6375	0.7043
Merge feature vector	Accuracy	**0.7182**	**0.9609**	**0.7911**	**0.8579**
	F1-score	**0.5567**	**0.9633**	**0.7631**	**0.8134**

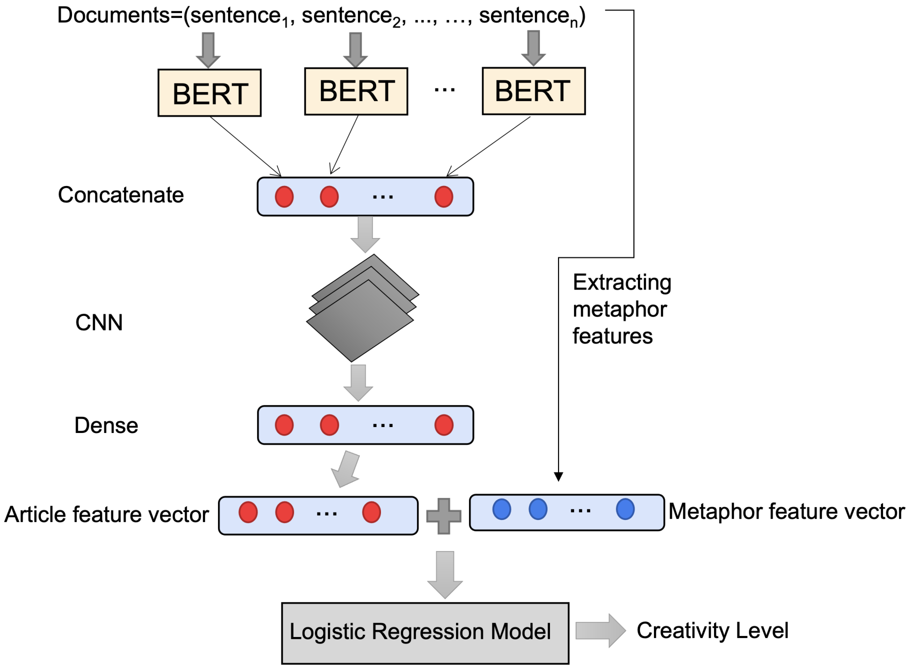
Documents=(sentence$_1$, sentence$_2$, ..., ..., sentence$_n$)
BERT
BERT
...
BERT
Concatenate
CNN
Dense
Extracting metaphor features
Article feature vector
Metaphor feature vector
Logistic Regression Model
Creativity Level

Figure 2: The architecture of our creativity assessment model.

feature is unavailable in the corresponding dataset. Values with superscript # are the chi-square statistic.

After extracting metaphor features in the datasets, we study the correlation coefficient between every feature and label using the point-biserial correlation coefficient, as most features are continuous variables, and the labels of creativity are binary variables. The coefficient is:

$$r_{pb} = \frac{\bar{X}_1 - \bar{X}_0}{s_X}\sqrt{\frac{N_0 N_1}{N(N-1)}}, \tag{1}$$

where $\bar{X}_0$ and $\bar{X}_1$ are the means of samples' features labeled as 0 and 1, respectively. N_0 and N_1 are the number of samples labeled as 0 and 1, respectively. N is the total number of samples, and s_X is the standard deviation of all samples' features (Table 2). Features with higher correlations may perform better in creativity classification in the student writing dataset. Most features have high correlations with creativity in the news and journal paper datasets.

4.2.2 Creativity Classification Method. The process of the writing creativity assessment model presented in this paper is shown in Figure 2. We first use the BERT model to convert the sentences in the document into a vector form, and then concatenate the sentence vectors into a matrix which contains the deep semantic information about the text. We input the obtained matrix into a convolutional neural network (CNN), and extract the metaphor feature of the text contained in the matrix through convolution and pooling operations. The dimension is reduced by a fully connected layer, and the obtained vector is used as the feature vector of the article. Finally, the metaphor feature vector is connected to the article feature vector, and the feature vector is classified by the logistic regression classifier to obtain the final classification result, which is the creative level of the article.

5 CREATIVITY ASSESSMENT EXPERIMENTS

5.1 Baseline

We consider the work of Ghosal et al. [8], which has the state-of-the-art performance at the creativity-related task, as our baseline. We use their method to classify documents' creativity levels, which has four sub-modules. The embedding module analyzes each sentence

Table 4: Performance of baseline and our method using all features.

Methods	Standard	Student Writing	News	Journal Paper	Webis-CPC-11
Our method	Accuracy	**0.7182**	**0.9609**	**0.7911**	**0.8579**
	F1-score	**0.5567**	**0.9633**	**0.7631**	**0.8134**
Baseline	Accuracy	0.6271	0.9331	0.6449	0.7422
	F1-score	0.4728	0.9518	0.6162	0.6867

Table 5: Performance of different metaphor identification models.

Methods	Metaphor	Student Writing	News	Journal Paper	Webis-CPC-11
$SIM - SG_I$	0.70	0.5033	0.9201	0.7181	0.7333
$SIM - SG_{I+O}$	0.73	0.5183	0.9321	0.7242	0.7483
$SIM - CBOW_I$	0.72	0.4731	0.9234	0.7234	0.7031
BERT+Bi-GRU+CRF	**0.84**	**0.5567**	**0.9633**	**0.7631**	**0.8134**

with a sentence encoder based on a bidirectional LSTM structure with max-pooling. The sentence encoder transforms sentences into fixed-size vectors. Then, the comparator module chooses the most similar source sentence b_{ij} to every sentence a_k in the target document according to the cosine similarity of sentence vectors. The aggregator module creates a relative sentence vector (RSV) corresponding to the target sentence:

$$RSV_k = [a_k, b_{ij}, |a_k - b_{ij}|, a_k * b_{ij}]. \tag{2}$$

It aggregates RSVs to get relative document vectors (RDVs) of target documents with dimension $N \times 4D$, where N is number of sentences in target documents and D represents sentence vector dimension. RDV is input to the CNN module that produces feature maps of implicit features. It obtains global features of the target document via a max-pooling softmax layer.

5.2 Experimental settings

In creativity classification, we perform a 10-fold cross validation on our datasets. The sentence vector dimension is set to 768. In order to unify the length of documents, we set the threshold to 50, and the part of the document with more than 50 sentences will be ignored. For documents with less than 50 sentences, we fill the matrix with a full-zero sequence of 768 dimensions. We use convolution kernels of sizes 2, 3, 4, and 5 for convolution operations, with a convolution layer window size of 150, and pooling operations using both max-pooling and average-pooling. The full connection layer dimensions are set to 100, 50, and 15, respectively, and the resulting document feature vector has a dimension of 15. After normalizing the five metaphorical features mentioned above, they are connected to the end of the article feature vector to form a new merge feature vector. Then, the merge feature vector is inputted into a logistic regression model, in which penalty is "l2", C is "1.0", the solver is "newton-cg", and other settings are default.

5.3 Experimental Results and Discussions

5.3.1 Performance on Different Datasets. We test the performance of creativity classification models, among which some consider metaphor features, and some do not consider metaphor features, on different datasets. The performance of each dataset is in Table 3. The highest accuracy comes on models with merge feature vector, followed by models with document feature vector, and metaphor feature vector.

On the student writing dataset, the combination of document feature vector and metaphor feature vector has the best performance on both accuracy (0.7182) and F1-score (0.5567). Thus, metaphor-based features can improve the results of creativity assessment. The prediction results on the news dataset are best. Table 3 shows that when document feature and metaphor feature are put together, both accuracy (0.9609) and F1-score (0.9633) out-perform document feature alone. Thus, metaphor features are very efficient at assessing creativity in writing and it in particular significantly improved the results. The results for journal papers and Webis-CPC-11 are similar. The prediction performance will be improved if the metaphor feature is added. Thus, metaphor features act as indicators of the assessment of creativity in writing.

5.3.2 Comparison with Baseline. In experiments on different datasets using the baseline method, we choose five creative and five non-creative documents to create RDVs. The number of sentences is 14, and the sentence vector is encoded by an open-source tool in a fixed size (4,096 dimensions). We adjust the size of filter windows (h) to 2,3,4 with 100 feature maps each and the number of training iterations at 100, since these documents are of different quality and length. We compare our method using all features with the baseline in four datasets (Table 4). Our model outperforms baseline in all the datasets, so the features and method we apply are useful and have strong interpretability.

5.3.3 Metaphor analysis. We also conduct an experiment to verify how effective the metaphor identification algorithm is; that is, how the performance of the creativity evaluation algorithm is dependent on the metaphor detection algorithm. In order to exclude our model's dependence on metaphor identification algorithms, we try to re-extract metaphor features using three metaphor recognition models with similar effects. Specifically, We utilize the three models $SIM - SG_I$, $SIM - SG_{I+O}$, $SIM - CBOW_I$ mentioned by Mao et al.[15]. Then we use these features in the creativity classification task. We test whether the results of these models differ when the features are different. The results show that the performance of creativity classification tasks varies with the variation of the performance of metaphor detection. The F1 scores are shown in Table 5. The second column is the results of the metaphor identification task. The last three columns are the results of the creativity identification task. The results show that our model does not depend on the metaphor identification model, but only on the metaphor itself.

6 CONCLUSION

This paper proposes a new metaphor identification-based writing creativity assessment model. The results have revealed that metaphorical expressions not only present vivid language, but also provide a cue to creativity. Our algorithm significantly outperforms the state-of-the-art method. To the best of our knowledge, our metaphor identification model has the most advanced recognition available, and we are the first to use automatic metaphor identification to assess writing creativity. Due to the scarcity of relevant work, our datasets with models and results may help with computational creativity and related problems.

Our study provides insights into and potential implications of educational utilities. For example, by integrating the use of automatic metaphor identification into the syllabus, educators could investigate novel pedagogical methods relating to creativity improvement. This creativity assessment approach using automatic metaphor identification inspires researchers working on automated scoring.

ETHICAL CONSIDERATIONS

We collect publicly available and widely used datasets and the collection process does not involve with privacy rights. For the creativity annotation, the salary for annotating each sample is determined by the average time and difficulty of the annotation to ensure that annotators are fairly compensated.

ACKNOWLEDGMENTS

This work is supported by National Natural Science Foundation of China under Grant No. 62076051.

REFERENCES

[1] Vahid Aryadoust, Li Ying Ng, and Hiroki Sayama. 2021. A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing* 38, 1 (2021), 6–40.

[2] Xiaomei Bai, Fuli Zhang, Jinzhou Li, Teng Guo, Abdul Aziz, Aijing Jin, and Feng Xia. 2021. Educational Big Data: Predictions, Applications and Challenges. *Big Data Res.* 26 (2021), 100270. https://doi.org/10.1016/j.bdr.2021.100270

[3] Majdi H. Beseiso, Omar A. Alzubi, and Hasan Rashaideh. 2021. A novel automated essay scoring approach for reliable higher educational assessments. *J. Comput. High. Educ.* 33, 3 (2021), 727–746. https://doi.org/10.1007/s12528-021-09283-1

[4] Xin Chen, Zhen Hai, Suge Wang, Deyu Li, Chao Wang, and Huanbo Luan. 2021. Metaphor identification: A contextual inconsistency based neural sequence labeling approach. *Neurocomputing* 428 (2021), 268–279. https://doi.org/10.1016/j.neucom.2020.12.010

[5] Minjin Choi, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. 2021. MelBERT: Metaphor Detection via Contextualized Late Interaction using Metaphorical Identification Theories. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (Eds.). Association for Computational Linguistics, 1763–1773. https://doi.org/10.18653/v1/2021.naacl-main.141

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/v1/n19-1423

[7] Rosie Dunford, Quanrong Su, and Ekraj Tamang. 2014. The pareto principle. (2014).

[8] Tirthankar Ghosal, Vignesh Edithal, Asif Ekbal, Pushpak Bhattacharyya, George Tsatsaronis, and Srinivasa Satya Sameer Kumar Chivukula. 2018. Novelty Goes Deep. A Deep Neural Solution To Document Level Novelty Detection. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, Emily M. Bender, Leon Derczynski, and Pierre Isabelle (Eds.). Association for Computational Linguistics, 2802–2813. https://aclanthology.org/C18-1237/

[9] José-Ángel González, Lluís-F. Hurtado, and Ferran Pla. 2021. TWilBert: Pre-trained deep bidirectional transformers for Spanish Twitter. *Neurocomputing* 426 (2021), 58–69. https://doi.org/10.1016/j.neucom.2020.09.078

[10] Saqib Ali Khan, Syed Muhammad Daniyal Khalid, Muhammad Ali Shahzad, and Faisal Shafait. 2020. Table Structure Extraction with Bi-directional Gated Recurrent Unit Networks. *CoRR* abs/2001.02501 (2020). arXiv:2001.02501 http://arxiv.org/abs/2001.02501

[11] Adam Kisvari, Zi Lin, and Xiaolei Liu. 2021. Wind power forecasting–A data-driven method along with gated recurrent neural network. *Renewable Energy* 163 (2021), 1895–1909.

[12] Vivekanandan Suresh Kumar and David Boulanger. 2021. Automated Essay Scoring and the Deep Learning Black Box: How Are Rubric Scores Determined? *Int. J. Artif. Intell. Educ.* 31, 3 (2021), 538–584. https://doi.org/10.1007/s40593-020-00211-5

[13] Jerry Chun-Wei Lin, Yinan Shao, Ji Zhang, and Unil Yun. 2020. Enhanced sequence labeling based on latent variable conditional random fields. *Neurocomputing* 403 (2020), 431–440. https://doi.org/10.1016/j.neucom.2020.04.102

[14] Jiawei Liu, Yang Xu, and Lingzhe Zhao. 2019. Automated Essay Scoring based on Two-Stage Learning. *CoRR* abs/1901.07744 (2019). arXiv:1901.07744 http://arxiv.org/abs/1901.07744

[15] Rui Mao, Chenghua Lin, and Frank Guerin. 2018. Word Embedding and WordNet Based Metaphor Identification and Interpretation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, Iryna Gurevych and Yusuke Miyao (Eds.). Association for Computational Linguistics, 1222–1231. https://doi.org/10.18653/v1/P18-1113

[16] Anita Milicevic, Sue Woolfe, Angela Blazely, Rhoshel Lenroot, and Stephen Sewell. 2020. Enhancing creativity through seven stages of transformation in a graduate level writing course—A mixed method study. *Thinking Skills and Creativity* 38 (2020), 100712. https://doi.org/10.1016/j.tsc.2020.100712

[17] Shangchao Min and Vahid Aryadoust. 2021. A systematic review of item response theory in language assessment: Implications for the dimensionality of language ability. *Studies in Educational Evaluation* 68 (2021), 100963.

[18] Arthur Neidlein, Philip Wiesenbach, and Katja Markert. 2020. An analysis of language models for metaphor recognition. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, Donia Scott, Núria Bel, and Chengqing Zong (Eds.). International Committee on Computational Linguistics, 3722–3736. https://doi.org/10.18653/v1/2020.coling-main.332

[19] Ciyuan Peng and Jason J. Jung. 2021. Interpretation of metaphors in Chinese poetry: Where did Li Bai place his emotions? *Digit. Scholarsh. Humanit.* 36, 2 (2021), 421–429. https://doi.org/10.1093/llc/fqaa016

[20] Ciyuan Peng, Dang-Thinh Vu, and Jason J. Jung. 2021. Knowledge graph-based metaphor representation for literature understanding. *Digit. Scholarsh. Humanit.* 36, 3 (2021), 698–711. https://doi.org/10.1093/llc/fqaa072

[21] Malay Pramanick, Ashim Gupta, and Pabitra Mitra. 2018. An LSTM-CRF Based Approach to Token-Level Metaphor Detection. In *Proceedings of the Workshop on Figurative Language Processing, Fig-Lang@NAACL-HLT 2018, New Orleans, Louisiana, 6 June 2018*, Beata Beigman Klebanov, Ekaterina Shutova, Patricia Lichtenstein, Smaranda Muresan, and Chee Wee Leong (Eds.). Association for Computational Linguistics, 67–75. https://doi.org/10.18653/v1/W18-0908

[22] Andrea Schiavio and Mathias Benedek. 2020. Dimensions of musical creativity. *Frontiers in Neuroscience* 14 (2020), 1208.

[23] Chuandong Su, Fumiyo Fukumoto, Xiaoxi Huang, Jiyi Li, Rongbo Wang, and Zhiqun Chen. 2020. DeepMet: A Reading Comprehension Paradigm for Token-level Metaphor Detection. In *Proceedings of the Second Workshop on Figurative Language Processing, Fig-Lang@ACL 2020, Online, July 9, 2020*, Beata Beigman Klebanov, Ekaterina Shutova, Patricia Lichtenstein, Smaranda Muresan, Chee Wee Leong, Anna Feldman, and Debanjan Ghosh (Eds.). Association for Computational Linguistics, 30–39. https://doi.org/10.18653/v1/2020.figlang-1.4

[24] Qimeng Yang, Long Yu, Shengwei Tian, and Jinmiao Song. 2021. Collaborative semantic representation network for metaphor detection. *Applied Soft Computing* (2021), 107911.

[25] Dongyu Zhang, Nan Shi, Ciyuan Peng, Abdul Aziz, Wenhong Zhao, and Feng Xia. 2021. MAM: A Metaphor-Based Approach for Mental Illness Detection. In *Computational Science - ICCS 2021 - 21st International Conference, Krakow, Poland, June 16-18, 2021, Proceedings, Part III (Lecture Notes in Computer Science, Vol. 12744)*, Maciej Paszynski, Dieter Kranzlmüller, Valeria V. Krzhizhanovskaya, Jack J. Dongarra, and Peter M. A. Sloot (Eds.). Springer, 570–583. https://doi.org/10.1007/978-3-030-77967-2_47

[26] Dongyu Zhang, Minghao Zhang, Teng Guo, Ciyuan Peng, Vidya Saikrishna, and Feng Xia. 2021. In Your Face: Sentiment Analysis of Metaphor with Facial Expressive Features. In *International Joint Conference on Neural Networks, IJCNN 2021, Shenzhen, China, July 18-22, 2021*. IEEE, 1–8. https://doi.org/10.1109/IJCNN52387.2021.9533972

[27] Dongyu Zhang, Minghao Zhang, Ciyuan Peng, Jason J. Jung, and Feng Xia. 2021. Metaphor research in the 21st century: A bibliographic analysis. *Comput. Sci. Inf. Syst.* 18, 1 (2021), 303–321. https://doi.org/10.2298/CSIS201109059Z