Off-topic English Essay Detection Model Based on Hybrid Semantic Space for Automated English Essay Scoring System

Guimin Huang, Jian Liu[a], Chunli Fan and Tingting Pan

School of Information and Communication Engineering, Guilin University of Electronic Technology,Guilin,China

Abstract. Aiming at the problem that the lack of accurate and efficient off-topic detection model for current Automated English Scoring System in China, an unsupervised off-topic essay detection model based on hybrid semantic space was proposed. Firstly, the essay and its essay prompt are respectively represented as noun phrases by using a neural-network dependency parser. Secondly, we introduce a method to construct a hybrid semantic space. Thirdly, we propose a method to represent the noun phrases of the essay and its prompt as vectors in hybrid semantic space and calculate the similarity between the essay and its prompt by using the noun phrase vectors of them. Finally, we propose a sort method to set the off-topic threshold so that the off-topic essays can be identified efficiently. The experimental results on four datasets totaling 5000 essays show that, compared to the previous off-topic essay detection models, the proposed model can detect off-topic essays with higher accuracy, and the accuracy rate over all essay data sets reaches 89.8%.

1 Introduction

Automated Essay Scoring(AES) system is an education software as using computer technology to evaluate and score the written essays[1], compared with manual scoring, it has the advantages of high efficiency and low cost. Baker[2] mentioned that it was important to limit the opportunity to submit uncooperative responses to education software. When a student enters a "good essay" that is unrelated to the essay topic, if there is no off-topic detection algorithm in the AES system, the AES system may give a higher score for the essay. Therefore, off-topic English essay detection algorithm is helpful to improve the fairness, robustness and accuracy of the AES system.

Off-topic detection algorithm is used to determine whether an essay is related to its topic. In AES system, there are two kinds of algorithms to detect off-topic essays. One kind of the algorithm belongs to the supervised algorithm, which requires topic specific training data to train the model in order to identify essays that are very different from the others on the same topic. The other kind of the algorithm belongs to the unsupervised algorithm which can identify the off-topic essay without using topic specific training data, it only uses the short prompt text on which the essay is supposed to have been written. In the actual situation, there are situations in which no topic specific training data are available for training. In addition, even model essays which are used to compare similarity with the essay text may not be sufficient sometimes. Therefore, the unsupervised off-topic essay detection algorithm has become the main research content of off-topic essay detection algorithm in recent years. The key of the unsupervised off-topic essay detection algorithm is to capture the similarity between the essay and its prompt. Inspired by the Term Frequency-Inverse Document Frequency (TF-IDF), Higgins et al.[3] proposed an off-topic essay detection method which used cosine similarity between TF-IDF vectors of an essay and its prompt to calculate the similarity between a prompt-essay pair. However, the TF-IDF vectors are not able to capture the semantic similarity between words such as “dog” and “canine”. On the basis of TF-IDF, Louis and Higgins[4] used WordNet to expand the words of short prompt with similar words to enable better comparison of essay text and its prompt. However, this method relies too much on artificial lexicon and may encounter some problems when words are not included in the lexicon. In order to further obtain the semantic similarity between words, some distributional word embedding techniques such as Mikolove et al.’s word2vec[5] and Pennington et al.’s GloVe[6] were proposed. On the basis of Mikolov, Rei and Cummins[7] proposed an improved algorithm to calculate the similarity of an essay and its prompt. The similarity algorithm extended the well-known Word2Vec embeddings by weighting them with TF-IDF to represent a sentence as a sentence vector, and then the cosine similarity between the sentence vectors can be used to get the similarity between sentences. By experimenting in a real essay data set, the results show that the method has strong robustness. However, the Word2Vec word embeddings always lack representation of relational knowledge. For example, it could not get the semantic correlation between “drink bear” and “car crash”. As we all know, English essay test always correlate with some

[a] Corresponding author: 239717061@qq.com

representation of relational knowledge. When the essay prompt is "The problem of drinking too much", if a student write "It may cause car crash", the existing algorithms will judge it unrelated to the prompt. In allusion to the deficiencies of the above existing model, we propose a hybrid semantic space based off-topic essay detection model which combine the distributional semantics and relational knowledge to enable better comparison of an essay text and its prompt in a hybrid semantic space.

In this paper, the off-topic essay detection model is described in Section 2. Section 3 introduces the training corpus of the model. Section 4 shows the experimental results on four data sets totaling 5000 essays.

2 Hybrid Semantic space based off-topic essay detection model

We design the off-topic essay detection model by the following steps: firstly, we extract noun phrases from essay and essay prompt; secondly, we construct a hybrid semantic space. Finally, we represent the noun phrases of the essay and the essay prompt as vectors in hybrid semantic space, and propose an algorithm to calculate the similarity values between the essay and the essay prompt.

2.1 Noun phrase extraction

The object that a sentence wants to express is usually represented in noun phrases. To enable better comparison of the essay text and its prompt, we extract noun phrases from them. In this paper, we use the neural-network dependency parser to parse the sentence of essay. The parser was proposed by Chen[8]. Figure 1 shows the parsing of a sentence in an essay. A sentence is parsed into a syntax analysis tree. Each leaf node in Figure 1 represents a syntactic component of a sentence. After parsing the sentence, we use regular expressions to extract noun phrases from syntax analysis tree.

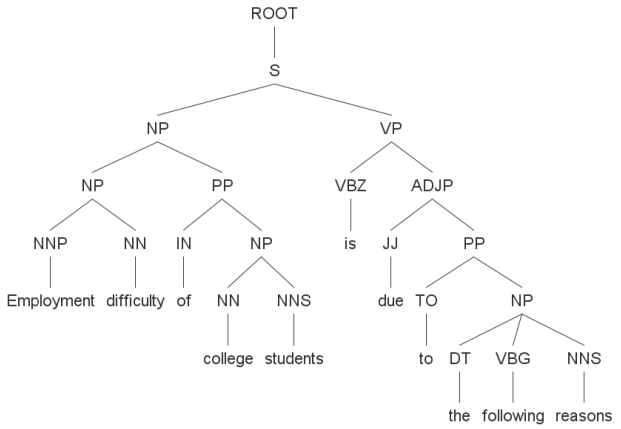

Figure 1. A sentence parsing example

2.2 Hybrid semantic space

Hybrid semantic space is a large word and phrase vector matrix which learns from both distributional semantics(such as word2vec and GloVe) and structured knowledge(such as ConceptNet[9] and PPDB[10]). To build a hybrid semantic space, Faruqui[11] proposed a method to "retrofitting" word2vec and GloVe word embeddings by using semantic lexicon. Based on the "retrofitting" method, Speer[12] proposed an effective hybrid semantic space called "ConceptNet Numberbatch". On the basis of Speer, in order to make the hybrid semantic space more suitable for representing the essay and its prompt, we construct a hybrid semantic space by using some synonyms and synonymous noun phrases that often appear in English essays to further retrofit the ConceptNet Numberbatch. The construction process of hybrid semantic space contains two cases. When the synonyms and synonymous noun phrases that we want to use to retrofit the ConceptNet Numberbatch exist in the ConceptNet Numberbatch, the purpose of the retrofitting is to make these synonyms and synonymous noun phrases set closer in our vector space. The retrofit steps are as follows:

Firstly, we represent ConceptNet Numberbatch as an initial matrix $\widehat{\mathbf{Q}}=\{\hat{\mathbf{q}}_1,\ldots\hat{\mathbf{q}}_n\}$, the semantic relations between the words in synonyms and synonymous noun phrases set as an undirected graph, secondly, we represent $\mathbf{Q}=\{\mathbf{q}_1,\ldots\mathbf{q}_n\}$ as a matrix to be infered, our propose is to make $\mathbf{q}_i$ close to its original values $\hat{\mathbf{q}}_j$ and their neighbors in the graph with edges E. Finally, based on the method of Faruqui[11], we can get the $\mathbf{Q}$ by minimizing the follow objective function:

$$\Psi(Q)=\sum_{i=1}^{n}\left[\alpha_i\left\|\mathbf{q}_i-\hat{\mathbf{q}}_j\right\|^2+\sum_{(i,j)\in E}\beta_{ij}\left\|\mathbf{q}_i-\mathbf{q}_j\right\|^2\right] \quad (1)$$

Where α and β values control the relative strenghs of associations.

When the synonyms and synonymous noun phrases that we want to use to retrofit the ConceptNet Numberbatch do not exist in the ConceptNet Numberbatch, the purpose of the retrofitting is to expand the hybrid semantic space with these synonyms and synonymous noun phrases and then make these synonyms and synonymous noun phrases set closer in our vector space. The expanded retrofitting steps are as follows:

Firstly, we merge the terms in ConceptNet with the synonyms and synonymous noun phrases set that we want to use to retrofit for transformation as a vocabulary, and let m be the size of it. Secondly, we define $\mathbf{S}$ is an $m\times m$ matrix which contains weighted values for terms that are known to be semantically related, and zero otherwise. The rows in the $\mathbf{S}$ add up to 1. Thirdly, we define $\mathbf{Q}^0$ is an $m\times n$ matrix, its rows are the original embeddings if available and let the rows be all zeros if the terms are outside the vocabulary of the original embeddings, then we define $\mathbf{A}$ is a diagonal matrix of weights in which $\mathbf{A}_{ii}$ is 1 if term i is in the original vocabulary, and 0 otherwise. Finally, based on the method of Speer[13], we can update $\mathbf{Q}$ iteratively so that the next interation of $\mathbf{Q}$ is a combination of its product with $\mathbf{S}$ and its weighted original state, followed by L_2 normalization of its non-zero rows:

$$\mathbf{Q}^{k+1}=normalize\left[\left(\mathbf{S}\mathbf{Q}^K+\mathbf{A}\mathbf{Q}^0\right)\left(\mathbf{E}+\mathbf{A}\right)^{-1}\right] \quad (2)$$

Where the $\mathbf{S}$ matrix relates to each term by the diagonal of it, and we find that the addition of 1 to the diagonal line

has a great effect on the convergence of the expanded retrofitting.

After retrofitting the hybrid semantic space, it can show more semantic relationships of words or phrases between essays.

2.3 Off-topic essay detection

Based on the hybrid semantic space, we can represent the words and phrases which exist in the hybrid semantic space as vectors. The hybrid semantic space are large enough and almost all the words used for English essay are included in it, but there are some phrases which are used in the essays and essay prompts are not included in it. So we use a simple but high performance method which was proposed by Arora[14] to get the phrase vector by computing the weighted average of the word vectors in the phrase and then remove the projections of the average vectors on their first principal component. On the basis of the above method, we propose a method to get the relationship between the essay and essay prompt in the hybrid semantic space, the main steps are as follows:

Firstly, we parse the essay in sentence and extract the noun phrases from each sentence, then represent the noun phrases as vectors in hybrid semantic space. Secondly, we extract the noun phrases from the essay prompt and represent the noun phrases of essay prompt as vectors in hybrid semantic space. Finally, we design an equation to calculate the relationship between the essay and essay prompt. The *Score(E,P)* indicates the relationship between the essay *E* and the essay prompt *P*.

$$Score(E,P)=\frac{1}{N}\sum_{i=1}^{N}\max_{k=1}^{m}\{sim(\mathbf{P}_{ij},\mathbf{Q}_k)\} \quad (3)$$

Where N is the total number of sentences on the essay, $\mathbf{P}_{ij}$ is the jth noun phrase vector in the sentence i of the essay and is of length 300. $\mathbf{Q}_k$ is the kth noun phrase vector of the essay prompt and is of length 300. $sim(\mathbf{P}_{ij},\mathbf{Q}_k)$ is the cosine similarity of $\mathbf{P}_{ij}$ and $\mathbf{Q}_k$. The value of the *Score(E,P)* is between 0 to 1.

In order to determine whether the essay under test is biased to other prompts compared with its own prompt, we construct an essay prompts set which contains 200 essay prompts from CET-4(College English Test 4), CET-6, and Ten-thousand English Compositions of Chinese learners(TECCL). When an essay is on-topic, it will be semantically similar to its prompt rather than other essay prompts. Therefore, we use the above similarity method to analyse whether an essay is off-topic or not, the main steps are as follows:

Firstly, we use the equation (3) to get the similarity value between the essay and its prompt. Secondly, we use the equation (3) to get the similarity values between the essay and all essay prompts of the essay prompts set. Finally, we sort these similarity values, when the value of the similarity between the essay and its prompt are in the top m, the essay is considered on-topic, otherwise the essay is considered to be off-topic. The ranking threshold m will be derived from the experimental part.

3 Hybrid semantic space retrofitting corpus

Our hybrid semantic space is based on Concept Numberbatch. ConceptNet Numberbatch is a semantic space, and its vocabulary is derived from word2vec, GloVe and the pruned ConceptNet graph. The word2vec vectors were trained on 100 billion words of Google news data set and are of length 300. The GloVe vectors were trained on 6 billion words from Wikipedia and English Gigaword and are of length 300. The ConceptNet 5.5 is a knowledge graph which include world knowledge from many different sources such as Open Mind Common Sense(OMCS) and information extracted from parsing Wiktionary.

On the basis of ConceptNet Numberbatch, we use some lexicons to retrofit it. The lexicons which were used to retrofit the ConceptNet Numberbatch include the Oxford Study Thesaurus, and the paraphrase database(PPDB) which is a semantic lexicon containing more than 220 million paraphrase pairs of English. To make the hybrid semantic space more suitable for representing essays, we extract synonymous noun phrases from International Corpus of Learner English(ICLE) and Ten-thousand English Compositions of Chinese learners(TECCL) to retrofit the hybrid semantic space. There are about 16000 essays written to over 1000 different essay prompts in ICLE and TECCL, and in total, we have extracted nearly 1000 sets of synonymous noun phrases to retrofit the hybrid semantic space.

4 Experiment

The datasets that we use to evaluate our off-topic essay detection model contain a total of 5000 student essays which are written to 25 different prompts or topics. The 5000 student essays consist of four essay sets: 500 essays drawn from CET-4, 500 essays drawn from CET-6, 1500 essays drawn from Chinese English Learner Corpus(CELC) and 2500 essays drawn from Kaggle competition data set. The first three data sets were written by Chinese students and the fourth essay data set is written by native English students. The off-topic essays of the datasets mainly include two different parts, one part of the off-topic essays are artificially judged as off-topic, the other part of the off-topic essays are essays which were randomly selected from other topics. And the essays in CET-4 set include 5 topics, 80 on-topic essays and 20 off-topic essays per topic. The essays in CET-6 set include 5 topics, 80 on-topic essays and 20 off-topic essays per topic. The essays in CLEC set include 10 topics, 120 on-topic essays and 30 off-topic essays per topic. The essays in Kaggle competition data set include 5 topics, 400 on-topic essays and 100 off-topic essays per topic. So the 5000 student essays contain a total of 4000 on-topic essays and 1000 off-topic essays.

We evaluate the performance of our off-topic essay detection model by the false positive rate(FPR), false negative rate(FNR) and accuracy rate. The false positive rate is the persentage of off-topic essays that have been incorrectly identified as on-topic; the false negative rate is

the pecentage of true on-topic essays that have been incorrectly identified as off-topic; the accuracy rate is the percentage of essays that have been correctly identified.

Our model needs to sort the values of similarity as described in section 2.3, therefore, the value of the threshold *m* should be obtained through the experiment. We set the value of *m* to 1-25, and conduct the off-topic essay detection experiment for 5000 student essays respectively, and then calculate the corresponding accuracy rate. When the ranking threshold *m* is 15, the accuracy of our off-detection model is the maximum of 89.80%. So in the following experiment, we set the value of *m* to be 15.

We take the TF-IDF and WordNet based off-topic essay detection method which was proposed by Louis and Higgins[5] as the benchmark method. Before the experiment, inspired by Louis and Higgins[5], we use the spelling correction method to correct the spelling in the essays, and then we conduct the off-topic essay detection experiments in above four essay sets. In the experiments, our model will compare with the benchmark method and Rei's word2vec based method[9], and we use FPR, FNR to evaluate the performance of three off-topic essay detection models. The experimental results are shown in Table 1.

Table 1. The experimental result of three methods on four data sets

Dataset	TF-IDF+WordNet		Word2Vec		Our model	
	FP%	FN%	FP%	FN%	FP%	FN%
CET-4	4.60	11.80	4.20	10.40	2.40	8.20
CET-6	5.00	11.40	5.40	10.00	2.65	8.40
CLEC	5.33	12.20	5.07	10.93	2.87	6.87
Kaggle	3.68	9.84	3.92	9.96	3.20	7.04
Total	4.40	10.90	4.44	10.30	2.96	7.24

According to the experimental results on four different data sets, we can find that the FPR and FNR of our off-topic detection model are lower than the other two models, especially for judging Chinese students' English essays, our model is better than the other two models. And the FPR over all data sets of our model is only about 2.96%, that means the probability that an off-topic essay is judged to be on-topic essay is very low. The probability of judging the on-topic essays as off-topic essays is 7.24%, which is relatively high. The reason is that the prompts in the essay prompts set of this model is relatively rich and comprehensive, and when the under test essay's prompt is short and contains less information, the essay to be test will be more similar to the prompts in the essay prompts set than its own prompt. Above all, the accuracy rate over all data sets of our model is 89.80%, and it can effectively detect whether the essay is off-topic or not.

5 Conclusion

This paper proposes an off-topic essay detection model by calculating the similarity value between the essay and the essay prompt in a hybrid semantic space. For improving the performance of our model, on the one hand, we extract the noun phrases from the essay and the essay prompt, which can effectively reduce the influence of the noise words on the off-topic analysis. On the other hand, we construct a hybrid semantic space which can represent both distributional semantics and structured knowledge, then we use some synonyms and synonymous noun phrases to further retrofit it and to make it more suitable for representing essays and essay prompts. Experimental results on multiple real data sets show that our off-topic model only needs essay prompt can identify whether the essay is off-topic or not effectively and accurately. Our model also significantly outperforms the previous off-topic detection models and will provide technical support for the AES system.

Acknowledgement

This work is supported by the National Natural Science Foundation of China (No. 61662012) as well as the Foundation of Key Laboratory of Cognitive Radio and Information Processing, Ministry of Education (Guilin University of Electronic Technology, No. CRKL150105).

References

1. Y. Attaly, J. Burstein. Automated essay scoring with e-rater® V. 2, **4(3)**, 1-31(2006)
2. R.S.J.d. Baker, A.M.J.B. De Carvalho, J. Raspat, V. Aleven, A.T. Corbett, K.R. Koedinger. *Educational software features that encourage and discourage "gaming the system"*, 475-482(2009)
3. D.Higgins, J. Burstein, Y. Attali. Identifying off-topic student essays without topic-specific training data, **12(2)**, 145-159(2006)
4. A. Louis, D. Higgins. *Off-topic essay detection using short prompt texts*, 92-95(2010)
5. T. Mikolov, K. Chen, G. Corrado, J. Dean. *Efficient estimation of word representations in vector space*, (2013)
6. J. Pennington, R. Socher, C.D. Manning. *GloVe: Global Vectors for Word Representation*, 1532-1543(2014)
7. M. Rei, R. Cummins. *Sentence Similarity Measures for Fine-Grained Estimation of Topical Relevance in Learner Essays*, 283-288(2016)
8. D. Chen, C.D. Manning. *A Fast Accurate Dependency Parser using Neural Networks*, 740-750 (2014)
9. H. Liu, P. Singh. ConceptNet — a practical commonsense reasoning toll-kit, **22(4)**, 211-226 (2004)
10. J. Ganitkevitch, B.V. Durme, C. Callison-Burch. *PPDB: The paraphrase database*, (2013)
11. M. Faruqui, J. Dodge, S.K. Jauhar, C.Dyer, E. Hovy, N.A. Smith. *Retrofitting Word Vectors to Semantic Lexicons*, (2015)
12. R. Speer, J. Chin, C. Havasi. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge, **31(2017)**, 4444-4451, (2017)

13. R. Speer, J. Chin. *An Ensemble Method to produce High-Quality Word Embeddings*, (2016)
14. S. Arora, L. Yingyu, M. Tengyu. *A Simple But Tough-to-Beat Baseline for Sentence Embeddings*, (2017)