Deep Learning Based Document Theme Analysis for Composition Generation

Jiahao Liu, Chengjie Sun, and Bing Qin

Harbin Institute of Technology, Harbin, China 150001
{jhliu, qinb}@ir.hit.edu.cn, sunchengjie@hit.edu.cn

Abstract. This paper puts forward theme analysis problem in order to automatically solve composition writing questions in Chinese college entrance examination. Theme analysis is to distillate the embedded semantic information from the given materials or documents. We proposes a hierarchical neural network framework to address this problem. Two deep learning based models under the proposed framework are presented. Besides, two transfer learning strategies based on the proposed deep learning models are tried to deal with the lack of large training data for composition theme analysis problems. Experimental results on two tag recommendation data sets show the effect of the proposed deep learning based theme analysis models. Also, we show the effect of the proposed model with transfer learning on a composition writing questions data set built by ourself.

Keywords: theme analysis, deep learning, transfer learning

1 Introduction

Automatically solving the material composition writing questions in a university's entrance examination like Gaokao [1] in China challenges natural language processing technology. Composition generation is a way to take up this challenge. Composition generation differs from text generation in that it needs to correctly analyze the theme firstly according to the specified materials. The target of text generation is to express the specified data in a natural language way. The specified data could be database records [2] or key words [3]. The key problem of text generation is language grounding [4], while the key problem for material composition generation is theme analysis.

How to analyze the theme from a given material? Because words are usually used to express the themes, theme analysis is related to key word extraction task [5]. Key words extraction methods can only output the words in the input text. However, the words expressing the theme of the material usually are not in the raw text and can capture the real semantic meaning behind the text. For example, the theme analysis results of the material in Fig 1 could be {爱岗敬业, 尽职尽责, 责任心} ({dedication, dutiful, responsibility}). Recently, [6] propose a deep keyphrase generation method, which attempts to capture the

作文阅读下面的材料，根据要求写一篇不少于800字的文章。（60分）
船主请一位修船工给自己的小船刷油漆。修船工刷漆的时候，发现船底有个小洞，就顺手给补了。过了些日子，船主来到他家里道谢，送上去一个大红包。修船工感到奇怪，说："您已经给过工钱了。"船主说："对，那是刷油漆的钱，这是补洞的报酬。"修船工说："哦，那只是顺手做的一件小事……"船主感激地说："当得知孩子们划船去海上之后，我才想起船底有洞这件事儿，绝望极了，觉得他们肯定回不来了。等到他们平安归来，我才明白是您救了他们。"
要求选好角度，确定立意，明确文体，自拟标题；不要脱离材料内容及含意的范围作文，不要套作，不得抄袭。

Read the following materials and write an essay of no less than 800 words. (60)
The owner of a boat asked a ship repairer to paint his boat. The repairer found that there was a small hole in the bottom of the boat during painting, he patched the hole by the way. A few days later, the owner came to thank him and gave him a big red envelope. The repairer felt surprised and said: "you have already paid." The owner said: "yes, that is for painting, this is the return to patch the hole." The repairer said: "Oh, just a little thing by the way....." The owner said gratefully: "when I heard that the children rowed out to the sea by the boat, I came to remember the bottom hole of the boat. I was so desperate and thought they couldn't come back. When they came back safely, I realized that you saved them."
Choose the right angle, determine the theme, make clear the genre, give the title; do not escape the scope of the content and meaning of the material, no plagiarism.

Fig. 1. An example of material composition question

deep semantic meaning of the content with a deep learning method. This work is also inspired by [6].

Another related task is tag recommendation [7]. Tag recommendation task has many similarities with this problem if we use tags to express the themes of materials. Both of them need to find tags to represent the meaning of a given text or materials. However one of the principle of making out the questions of Gaokao is to avoid similarity with history questions. So the successful collaborative filtering approaches in tag recommendation are not suitable for theme analysis for material composition question. Because themes are the distillation of materials, the theme analysis methods should have the ability to understand the materials.

Although there is still lack of clear explanation for the mechanism of deep learning, it does show the potential when dealing with semantic representation learning [8] and semantic reasoning [9]. Due to the promising of deep learning methods in natural language processing, deep learning based methods for document theme analysis required by composition generation are proposed in this work.

Deep learning models usually need large annotated training data to achieve good performance due to the numerous parameters in the models. However, no large annotated training data for theme analysis of material composition question is provided currently. One possible solution is to involve transfer learning [10] and some annotated training data for similarity tasks such as tag recommenda-

tion. Fortunately, the annotated training data for tag recommendation can be collected easily from some big social media websites such as Douban[1] and Zhihu[2]. Transfer learning for deep learning is also a hot research topic recently [11, 12]. We try several transfer learning strategies based on the proposed deep learning models in this work to prompt the performance of theme analysis for material composition question.

2 Problem Definition

Most of the composition writing questions in Gaokao are material compositions. In a material composition question, a short essay is given and the students are required to write a composition based on the theme embedded in given material as shown in the example of Fig 1. Theme analysis is the key step in the whole procedure of composition generation. It will be fail in this question if the theme is wrongly analyzed.

Theme analysis for material composition can be defined as following: given a short essay D, the target of theme analysis is to find a function F, which can map D to a word set $T = \{w_1, w_2, \cdots, w_n\}$. T represents the theme of D and can be used as the clue and input for composition generation.

Given the example in Fig 1, the output T of theme analysis function F should be { 爱岗敬业, 尽职尽责, 责任心} ({dedication, dutiful, responsibility}). The words in T are the sublimation of the given essay, which can not be obtained through literal comprehension.

There are 3-fold challenges for theme analysis of material composition:

- lack of large annotated material-theme pairs training data.
- theme is the distillation of materials, not the surface expression of that.
- the expression of theme needs to be suitable for following procedure of composition Writing.

3 Method

In this work, a hierarchical neural network framework is proposed to learn the semantic representation V_{doc} of the give short essay D in material composition writing questions. With this representation, a predictor can be trained to output the confidence score $\delta(w_i|V_{doc})$ for each candidate theme word w_i. The theme analysis results for a material composition writing question consist of the words with the top N confidence score. N could be defined according to the requirements of applications. A theme word vocabulary T could be built in previous.

Two models are presented under this framework in the following. One is based on Gated Recurrent Unit (GRU) [13], named GRU-GRU model; The other is based on Convolutional Neural Networks (CNN) [14] and GRU, named CNN-GRU Model.

[1] https://www.douban.com/
[2] https://www.zhihu.com/

3.1 GRU-GRU Model

The GRU-GRU model architecture is shown in Fig 2. In Fig 2, the bottom two recurrent neural network (RNN) parts encode the input text into semantic representation. The unit of the RNN layers is Gated Recurrent Unit. The word embedding (word vector) of each word in D is taken as the input of the network.

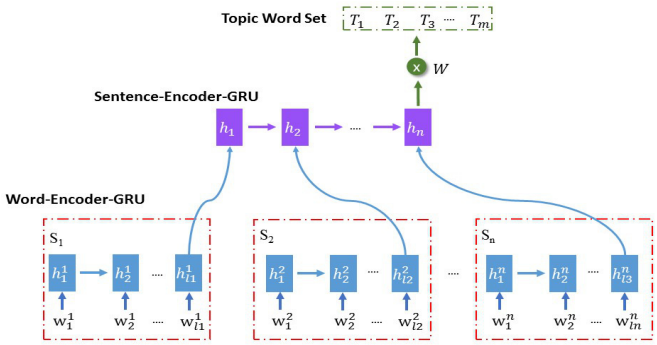

Fig. 2. GRU-GRU Model Architecture for Theme Analysis

The bottom part is a RNN word-sentence encoder. A "$\langle \backslash s \rangle$" is used to denote the end of a sentence. When encountering "$\langle \backslash s \rangle$", the hidden layer values of current RNN are taken as the sentence vector V_{s_i} of current sentence s_i. In this way, we can get $V_{s_1}, V_{s_2}, \cdots, V_{s_n}$ for a D with n sentences. The middle part is a RNN sentence-document encoder which takes V_{s_i} of each sentence s_i as input and outputs the semantic representation V_{doc} of D . The top part is a two-layer neural network. It takes document vector V_{doc} as input and output the confidence score δ_{w_i} of being theme word for each word w_i in T. δ_{w_i} is calculated by Eq 1, Φ_{w_i} is a row of matrix W in Fig 2. The size of matrix W is $|T| \times |V_{doc}|$.

$$\delta_{w_i} = Sigmoid(\Phi_{w_i} \cdot V_{doc}) \tag{1}$$

For a document D, the loss function of the network is defined as Eq 2. In Eq 2, T_D is the theme words set for D; M is the size of theme vocabulary T.

$$L = \sum_{i=1}^{M} [\sum_{w_i \in T_D} \log \delta_{w_i} + \sum_{w_i \notin T_D} \log(1 - \delta_{w_i})] \tag{2}$$

3.2 CNN-GRU Model

Convolutional Neural Networks (CNN) could better capture the local feature and has better performance when leaning the sentence semantic representation [15]. So, we propose CNN-GRU model by replacing the bottom layer in GRU-GRU model with CNN layer as shown in Fig 3. Other parts of the model are same as what in GRU-GRU model.

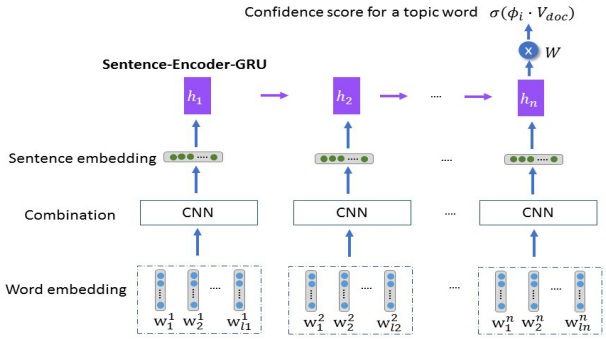

Fig. 3. CNN-GRU Model Architecture for Theme Analysis

3.3 Transfer Learning Strategies

In this section, two transfer learning strategies are proposed to overcome the shortage of theme analysis training data problem and boost the performance of theme analysis. Both of them are based on GRU-GRU model.

Feature Representation Based Transfer Learning In GRU-GRU model, V_{doc} can be considered as a document representation. Inspired by [16], we propose a feature representation based transfer learning strategy, which train the GRU-GRU model with source domain training data and re-train the top part of Fig 2 with target domain training data. Training data of source domain can be very large, so we can get a good document representation. Based on this transfered representation, better classifiers can be obtained by the small training data in the target domain.

Fine-tuning Based Transfer Learning In the proposed feature representation based transfer learning, only the parameters in the top part of Fig 2 are

modified by the target domain training data, which requires source domain and target domain have large similarities. While in fine-tuning based transfer learning, we first train the GRU-GRU model with source domain training data and then fine-tuning the whole network with the training data in target domain. In this way, all the parameters of GRU-GRU model learned from source domain will be adjusted to the target domain. When the differences between source domain and the target domain are large, this strategy may more suitable.

4 Data Set

Three data sets are used in this work: Composition, Zhihu and Douban. Table 1 shows samples of the three data set.

"Composition" data set is built by ourselves, which contains 1515 material composition problem and are annotated with theme-material pairs. The themes are expressed by words.

Zhihu and Douban are collected from two social media web sites: Zhihu and Douban. Zhihu data set are built by ourselves and 50,000 documents with their corresponding tags are downloaded from Zhihu Website. Douban data set came from Si [17] and Liu [18]. We compare our work with them on the same data set. The reasons why we use these two data sets are: 1) the shortage of large annotated material-theme pairs data set; 2) tags are given by users for numerous documents in the two websites, which could be considered as document theme words; 3) easy to compare with other works.

Data Set	**Document**	**Tags**
Composition	滑雪是一种很好的运动项目，穿越林海雪原，飞速行进在都市中无缘得见的皑皑大地上，体会从山坡上急速滑降时那种风驰电掣般的感觉，真是无限乐趣在其中。但滑雪者都清楚地知道，要想轻松愉快地顺着山坡往下滑行，就必须先背负器材、一步步辛苦地登上山顶。	**努力拼搏成功磨练困难克服困难**
Zhihu	和朋友一起创办了一个补习班，可朋友现在很不用心，已经很多家长由于她的原因选择让孩子离开我们的补习班了。和她谈了几次，可还是不用心。因为是当初创办时的费用是一人一半的，而且没约定任何事情，现在她这样我很为难，想让她退出，哪怕我损失些钱也好，我该怎么和她说呢?	**创业，合伙人，责任心**
Douban	本书作者佩珀·怀特曾就读于MIT的机械工程系，他凭借睿智的眼光、情文并茂的描写，对自己早年在MIT辛酸的求学、生活和创业经历进行了全景式的回放，以"身在其中"的方式展示出MIT的独特风貌，带给读者关于人文精神的深刻反思。在与作者共同缅怀这段黄金岁月的过程中，相信每一个对MIT心存向往的人，都能够身临其境，切实体验到在MIT学习、生活的每一个平实的日子。	**励志，我向往的学校，教育，传记，研究生，大学，MIT，经典，留学**

Table 1. Samples of Zhihu and Douban Data Set

5 Experimental Results

5.1 Experimental Settings

The experiments are designed for two purposes: 1) to show the theme analysis abilities of the two proposed deep learning models; 2) to verify the effect of the proposed transfer learning strategies.

Settings for deep learning based theme analysis 10,000 and 5,000 samples are randomly chosen from Zhihu data set as training data and test data respectively. The vocabulary T is built by collecting all the tag words in the training data. The embeddings of words are trained using the training data by word2vec[3] tools. The dimension of word vector is 200. In GRU-GRU model, the recurrent hidden layer of the word layer GRU contains 200 hidden units while sentences layer GRU contains 500. As to CNN-GRU model, we use three convolutional filters whose widths are 1, 2 and 3 to encode semantics of unigrams, bigrams and trigrams in a sentence. And the number of filters is 200, while the parameters of sentences layer GRU are the same as GRU-GRU model. Parameters of our model are randomly initialized over a uniform distribution with support [-0.01, 0.01]. The model is trained with the AdaDelta[19] algorithm.

The Top3 theme words given by the proposed models are used as the theme analysis results.

Precision, recall and F1-measure are taken as the evaluation criteria. The final evaluation scores are computed by micro-averaging (i.e. averaging on resources of test set). The tags given by users in the test data are taken as gold standard.

Settings for transfer learning The Zhihu data set and Composition data set are taken as source domain and target domain respectively. The detail information about the data used in transfer learning experiments are shown in Table 2.

Data set	**♯training**	**♯test**	**♯ candidate tags**
Composition (Target domain)	1415	100	694
Zhihu (Source domain)	10,000	5,000	5000

Table 2. Data settings for transfer learning

5.2 Experimental Results

Table 3 shows the results of the "GRU+GRU" and "CNN+GRU" models.

[3] https://code.google.com/archive/p/word2vec/

Method	Precision	Recall	F1-measure
GRU+GRU	0.2762	0.3173	0.2766
CNN+GRU	0.2828	0.3247	0.2828

Table 3. Experimental results on Zhihu Data Set

In order to better evaluate the performance of the proposed models, we compare their performances with TAM [17] and WTM [18] on Douban data set. With 49,050 documents with their corresponding tags from Douban as training data and 12,132 as test data, the comparison results are shown in Table 4.

Method	Precision	Recall	F1-measure
TAM	0.2971	0.3230	0.2676
WTM	0.3498	0.4182	0.3311
GRU+GRU	0.3680	0.4052	0.3337
CNN+GRU	0.3835	0.4213	0.3480

Table 4. Experimental results on Douban Data Set

Due to no titles are given for essays in material composition writing questions, our models don't deal with the title of a document. That's the reason why we didn't compare the results with title information in [17] and [18] .

From Table 4, it is obvious that two deep learning based models have better performance. The results indicate that deep learning based methods have better ability to understand the semantic of the documents than previous methods. Also, "CNN+GRU" model outperforms "GRU+GRU" model consistently on two data sets, which shows that CNN can use local information to obtain better sentence representation. Samples in Zhihu data set are more similar with material composition questions because most tags for a document can be found in the document in Douban data set. That's also the reason for the higher performance in Douban data set.

Table 5 shows the results of theme analysis on Composition data set with different methods. P@5 is used as the evaluation criteria. GRU+GRU model can only achieve 0.078 when we directly use 1415 training samples chosen from composition data set. The poor performance is largely due to the small number of training data compared with the results in Table 4. Two transfer learning methods can greatly boost the performance from 0.078 to more than 0.3. So transfer learning based on the deep learning model is a promising way to deal with theme analysis for material composition generation.

Method	P@5
GRU+GRU	0.078
Feature Representation	0.324
Fine-tuning	0.341

Table 5. Experimental results on Composition Data Set

6 Conclusion

The first step of automatic composition generation for material composition questions in Gaokao is to identify the theme of the given materials, which is even a big challenge for most high school students. This work proposes a deep learning framework to solve this problem. The contributions of this work lie in: 1)put forward theme analysis problem for material composition questions in Gaokao; 2) present two deep learning based methods to solve theme analysis problem and Show the potential of deep learning based theme analysis methods with two social media data sets; 3) propose transfer learning strategies to make the deep learning models trained on social media data set can be used to analyze material composition question data.

Acknowledgment

We would like to thank the anonymous reviewers for their thorough reviewing and proposing thoughtful comments to improve our paper. This work was supported by the National 863 Leading Technology Research Project via grant 2015AA015407, Key Projects of National Natural Science Foundation of China via grant 61632011.

References

1. Cheng, G., Zhu, W., Wang, Z., Chen, J., Qu, Y.: Taking up the gaokao challenge: An information retrieval approach. In: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016. (2016) 2479–2485
2. Konstas, I., Lapata, M.: A global model for concept-to-text generation. J. Artif. Intell. Res. (JAIR) **48** (2013) 305–346
3. Uchimoto, K., Sekine, S., Isahara, H.: Text generation from keywords. In: 19th International Conference on Computational Linguistics, COLING 2002, Howard International House and Academia Sinica, Taipei, Taiwan, August 24 - September 1, 2002. (2002)
4. Liang, P., Jordan, M.I., Klein, D.: Learning semantic correspondences with less supervision. In: ACL 2009, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2-7 August 2009, Singapore. (2009) 91–99

5. Mihalcea, R., Tarau, P.: Textrank: Bringing order into text. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing , EMNLP 2004, A meeting of SIGDAT, a Special Interest Group of the ACL, held in conjunction with ACL 2004, 25-26 July 2004, Barcelona, Spain. (2004) 404–411
6. Meng, R., Zhao, S., Han, S., He, D., Brusilovsky, P., Chi, Y.: Deep keyphrase generation. In: ACL17. (2017)
7. Sigurbjörnsson, B., van Zwol, R.: Flickr tag recommendation based on collective knowledge. In: Proceedings of the 17th International Conference on World Wide Web, WWW 2008, Beijing, China, April 21-25, 2008. (2008) 327–336
8. Bengio, Y., Courville, A.C., Vincent, P.: Representation learning: A review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. **35** (2013) 1798–1828
9. Schmidhuber, J.: Deep learning in neural networks: An overview. CoRR **abs/1404.7828** (2015)
10. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Trans. Knowl. Data Eng. **22**(10) (2010) 1345–1359
11. Zhuang, F., Cheng, X., Luo, P., Pan, S.J., He, Q.: Supervised representation learning: Transfer learning with deep autoencoders. In: Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015. (2015) 4119–4125
12. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In: Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada. (2014) 3320–3328
13. Cho, K., van Merrienboer, B., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL. (2014) 1724–1734
14. Kalchbrenner, N., Grefenstette, E., Blunsom, P.: A convolutional neural network for modelling sentences. In: ACL. (2014)
15. Kim, Y.: Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar,. (2014) 1746–1751
16. Razavian, A.S., Azizpour, H., Sullivan, J., Carlsson, S.: CNN features off-the-shelf: An astounding baseline for recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2014, Columbus, OH, USA, June 23-28, 2014. (2014) 512–519
17. Si, X., Liu, Z., Sun, M.: Modeling social annotations via latent reason identification. IEEE Intelligent Systems **25**(6) (2010) 42–49
18. Liu, Z., Chen, X., Sun, M.: A simple word trigger method for social tag suggestion. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL. (2011) 1577–1588
19. Zeiler, M.D.: ADADELTA: an adaptive learning rate method. CoRR **abs/1212.5701** (2012)