Leveraging Diverse Lexical Chains to Construct Essays for Chinese College Entrance Examination

Liunian Li Xiaojun Wan Jin-ge Yao Siming Yan
Institute of Computer Science and Technology, Peking University, Beijing 100871, China
The MOE Key Laboratory of Computational Linguistics, Peking University
{liliunian, wanxiaojun, yaojinge, dantes}@pku.edu.cn

Abstract

In this work we study the challenging task of automatically constructing essays for Chinese college entrance examination where the topic is specified in advance. We explore a sentence extraction framework based on diversified lexical chains to capture coherence and richness. Experimental analysis shows the effectiveness of our approach and reveals the importance of information richness in essay writing.

1 Introduction

Chinese National Higher Education Entrance Examination, a.k.a. Gaokao ("高考") in Chinese, has a similar format to the American SAT, except that it lasts more than twice as long. The nine-hour test is offered just once a year and is the sole determinant for admission to virtually all Chinese colleges and universities. It emphasizes several subjects including math and science, but also measures knowledge of written Chinese and English. It includes various types of questions such as multiple-choice questions, short-answer questions and essays. In this work, we focus on the Chinese essay writing questions, typically in the format of writing a topically rich but coherent essay when specified a topical word or title. Developing a system that can construct essays in the context of exams is not for mimicking or surpassing human writing, but to provide analytical assistance for students and high school teachers to improve essay writing. The task is challenging as it requires analyzing of the given topic and the ability to organize coherent descriptions in sentences and paragraphs, while the content should cover rich aspects and discussions but still conforms to the given topic. As a preliminary study we only explore sentence extraction to get a sense of how well automatic approaches could achieve when evaluated by professional evaluators.

We explore an approach based on lexical chains, i.e., sequences of words containing a series of conceptually related words in a discourse. Lexical chains could be useful to assist analyzing topical coherence and we will show their effectiveness in essay construction. For a given topic, we first retrieve a few topically relevant documents, from which the lexical chains will be built. Each lexical chain corresponds to a subtopic. We would like to have each subtopic representative, while the overall subtopics are diverse enough to cover as many topically related aspects as possible. We leverage a diversified ranking algorithm to calculate the importance weights for different lexical chains, then form the essay by selecting and organizing sentences to cover the important chains.

In this paper we provide a focused study on a specific topic. Our contributions can be summarized as follows:

- To the best of our knowledge, we are the first study Chinese essay generation for Chinese college entrance exams. We utilize sentence extraction as a viable step towards essay generation for exams.
- Considering the nature of the problem, we propose a framework based on diversified selection of lexical chains, to cover rich and diverse aspects of the given topical word.
- Manual evaluation from experienced high school teachers shows the feasibility for automatic essay generation, as well as the effectiveness and the potential of our approach, revealing the importance of information richness and diversity for essay writing in the context of Chinese college entrance exams.

Proceedings of the The 8th International Joint Conference on Natural Language Processing, pages 355–360,
Taipei, Taiwan, November 27 – December 1, 2017 ©2017 AFNLP

2 Approach

In this preliminary study, we directly use sentences from a large source corpus to construct the final essay. This can be treated as a viable intermediate step towards full generation. We design a framework that consists of article retrieval, lexical chain construction, sentence extraction and information ordering. Note that in this study the specified topic for essay writing is in the form of one single topical word while our approach relies on vectorial representations. Nevertheless, our solution can naturally generalize to sentential inputs, since we could use sentence embedding models (e.g. RNNs, skip-thought vectors) to get a vectorized representaion.

2.1 Article Retrieval

In this study we find that a simple retrieval model based on latent semantic analysis (Deerwester et al., 1990, LSA) works relatively well. Specifically, we get semantic vector representations for each document and the given topical word by performing singular value decomposition on the term-document matrix, where each element corresponds to the tf-idf value for a particular word in the document. Articles with the highest similarity scores with the given topic word will be retrieved for the next steps. We also tried an even simpler approach to directly use averaged word vectors to represent a document and search for the documents that lead to the largest cosine similarity values with the given topical word. This approach turns out to prefer shorter articles and yields less accurate performance compared with LSA.

2.2 Building Lexical Chains

The main task studied in this paper is to construct an essay for a specified topic word. A lexical chain (Morris and Hirst, 1991) is a sequence of words that consists of a series of conceptually related words in a discourse. A lexical chain can be used to model topical contexts as well as text coherence. In our study, we adapt the calculation method used by Barzilay and Elhadad (1999) for our purpose. Due to the shortage of high-coverage thesaurus, we utilize vector semantics to capture lexical relationships, and find that pairwise similarities defined by word vectors can be used to build reliable lexical chains while being more flexible.

We treat each retrieved article as a list of words. For the purpose of this study, we only consider adjectives and nouns when constructing the chains. Since in Chinese articles, most frequently used adverbs (such as “那么”-*so*) and verbs (such as “成为”-*becomes*, “有”-*exist*) are used for syntactic integrity and do not contain topically related information.

Given an article, we take out each word one by one and check whether and where it should be placed. If this word has not been included in any chain, we treat it as a candidate and traverse all current chains to calculate the similarity between the candidate and the chain. The candidate word will be attached to the chain with the highest similarity value if it surpasses a threshold. Here the similarity between a candidate word w and a chain C is defined as the similarity between w and the last word in C. [1]

2.3 Importance Estimation for Chains

Not all of the constructed lexical chains should be used for further processing, as some may not cover important aspects of the given topic. There could be many ways for estimating the importance of different lexical chains. In this work we model the problem using graph-based ranking, treating each candidate chain as a node in a graph. The edge weight is assigned to be pairwise similarity between two chains C_1 and C_2, defined as $cos(\frac{1}{\#C_1}\sum_{w\in C_1} vec(w),)$, i.e. the similarity between their average word vectors.

In order to cover more aspects about the topic and to avoid redundancy, we would like to assign high important scores on more diverse chains. Therefore we utilize DivRank (Mei et al., 2010), a well-known diversified ranking algorithm, for calculating the ranking scores for different chains. Specifically, we utilize the pointwise variant of DivRank. At time T, the transition probability from node u to node v is defined as follows:

$$p_T(u,v) = (1-\lambda)p^*(v) + \lambda\frac{p_0(u,v)p_T(v)}{D_T(u)} \quad (1)$$

where $p^*(v)$ is a distribution which represents the prior preference of visiting vertex v, $p_0(u,v)$ is the initialized transition matrix estimated from a regular time-homogenous random walk, and

[1] We find a more straightforward definition $\max_{t\in C} sim(w,t)$ less effective as it encourages a rich-gets-richer effect which leads to long chains but incoherent thematic meanings.

$D_T(u) = \sum_{v \in V} p_0(u,v) p_T(v)$ is a normalizing factor for the second term. The weights for each node are initialized to be the cosine similarity between the corresponding chain and the topic, which can in some sense reflect the prior relevance. After the algorithm converges, i.e. $p_T(v) \approx p_{T-1}(u) p_{T-1}(u,v)$, we can select the top-ranking chains as subtopics for essay construction according to the values of $p_T(v)$.

2.4 Sentence Selection

We now have our subtopics, i.e. important lexical chains, prepared. The next step is to select sentences to cover those subtopics. For each candidate sentence, we use the average vector of nouns and adjectives as its vectorial representation, denoted as $\mathbf{s}$. Let $\mathbf{c}$ and $\mathbf{t}$ be the average word vector of the current lexical chain and the vector of the topical word, respectively. We define the weight for sentence $\mathbf{s}$ to be a linear combination of chain similarity and topical word similarity:

$$weight(\mathbf{s}) = \frac{\cos(\mathbf{s}, \mathbf{t}) + \rho \cdot \cos(\mathbf{s}, \mathbf{c})}{1 + \rho} \qquad (2)$$

We empirically set the ratio parameter ρ to be 0.8 and observe that the selected sentences can cover the subtopic well while being coherent with the topical word.

An essay in Chinese college entrance exams normally has a length between 800 and 1,200 Chinese characters. The most typical essays contain around 1,000 characters. We find that taking seven top-ranking subtopics (chains) can lead to a good coverage of the given topic, and selecting four sentences for each subtopic to form a paragraph will make the overall length just around 1,000. Therefore we limit the numbers for subtopic and sentence selection as such.

2.5 Sentence Ordering

After selecting sentences for each subtopic, we need to order them to form paragraphs and order the paragraphs to construct the full essay. A straightforward way is to greedily select elements based on similarity between each candidate and the previously selected sentence or paragraph. Preliminary experiments suggest that most elements share high similarity values due to our selection criteria in previous steps, causing simple greedy selection strategy to fail. Therefore, we consider a different method: ordering sentences and paragraphs according to its position in the original article. The position of each candidate can be represented as a rational number, dividing the current position number by the total number of sentences in that paragraph. We find readability within a paragraph largely increased after such strategy for sentence ordering. The intuition is that sentences in the front or at the end typically contain more general discussion while sentences in the middle tend to describe specific details or expansive contents. For paragraph ordering, we take a similar approach by calculating relative positions in the original document.

3 Experimental Study

3.1 Settings

3.1.1 Data

Since we study approaches based on sentence extraction in this preliminary work, We collected a source corpus that contains 800 articles in Chinese, with the overall size around 800,000 Chinese characters. The source corpus consists of essays written by various authors, discussing relatively diverse topics. [2]

Since the similarity calculations in our framework involves vectorial representations for each word, we trained 300 dimensional GloVe vectors (Pennington et al., 2014) on the Chinese Gigaword corpus (Graff and Chen, 2005). We used the Stanford Chinese Segmenter for word segmentation (Tseng et al., 2005).

For evaluation, 10 topics which have once appeared in previous Chinese college entrance exams will be provided for all the experimented essay construction systems. We have manually checked that there indeed exists several sentences in the source corpus that are relevant to the given topic. A good system should detect such sentences and use them to generate the essay that well responds to the specified topic.

3.1.2 Evaluation

Given a topical word, every student will write a completely different essay. The diversity of possible essays makes automatic evaluation metrics that count on content overlaps impossible since system outputs can then only be compared with rather limited references. Therefore we leave proper design of automatic metrics as future work and only

[2] To promote related experimental and educational studies, we have attached the corpus in the supplementary materials which could be found in the ACL Anthology.

perform manual evaluation in this study.

We conduct manual ratings on a few important metrics (in a 1-10 rating system, the higher the better) in generic generation systems that also should be emphasized in this study, including

- **Topical consistency (const.)**: on how much is the output consistent with the given topic.
- **Overall readability (read.)**: overall readability of the essay in terms of text coherence.
- **Content diversity (div.)**: whether the essay covering multiple aspects of the topic or just repeating the same argument.

For the purpose of this study, we also evaluate the output essays using the evaluation criteria for the Chinese college entrance exams. The total rating score is 60 points, assigned for the *basic level* and the *advanced level* respectively. The basic level (40 points in total) considers whether the most basic requirements have been fulfilled, such as whether the essay conforms to the given topic, describing with structural integrity and using correct punctuation. The advanced level (20 points in total) measures how much depth, richness, literary grace and novelty there exist in the content of the essay. We ask 10 high school teachers who are experienced in such essay evaluation settings to conduct manual scoring. Note that in evaluations of the exams, the above points will not be strictly assigned one by one, only an overall score will be seen. We conform with this scoring approach in this study.

3.1.3 Baselines

To verify the effectiveness of our proposed approach, we compare with two baseline systems: The system (**Baseline1**) that utilizes topically related words and clusters rather than our proposed diversified lexical chains for subtopic representation, and a more straightforward baseline (**Baseline2**) that select sentences which have the most similar vectorial representations with the given topical word. The former baseline can be treated as a reimplementation of the very recent Chinese essay generating system proposed by Qin et al. (2015). All systems are evaluated on the given 10 topics, producing 30 essays in total.

3.2 Results

Table 1 lists the manual rating scores (average and standard deviation [3] of the scores from the 10 high school teachers) for the outputs from different systems. The differences between systems are statistically significant in Bonferroni adjusted pairwise-t testing with $p < 0.01$. We can see that our proposed framework outperforms the baseline systems in all evaluation criteria. We also provide the example outputs in the appendix.

	Baseline1	Baseline2	Proposed
Basic (40)	29.42±4.43	32.75±1.84	34.82±1.50
Adv (20)	11.65±2.53	13.1±1.79	14.97±1.23
Score (60)	41.07±6.22	45.85±3.51	49.78±2.65
const. (10)	5.38±0.98	6.47±0.84	6.90±0.76
read. (10)	5.23±0.81	6.27±0.75	6.78±0.68
div. (10)	5.15±0.74	5.8±0.87	6.92±0.75

Table 1: Evaluation results for different systems. Each cell contains the average and standard deviation of the ten scores assigned to an output.

3.3 Discussion

Here we provide some qualitative analysis for our use of lexical chains and diversified importance ranking. Given the topic word “挫折”(*setback*), we can find 40 chains in total. the top-ranking chains from diversified ranking are displayed in Figure 1a, along with the top-ranking subchains produced by the PageRank algorithm (Page et al., 1999) in Figure 1b. We can observe that chains in (1a) contain direct explanations as well as consequential attitudes and related association of words, while most chains in (1b) only cover the literal meaning of the word “挫折”(*setback*), without extensions in depth.

We also made some statistics to make sure that all systems are not directly using the original source documents. All systems produced essays using sentences from multiple articles between 9 and 21, with the overlapping proportion for single source document no more than 15%. This is intuitively a side evidence on that if a student wants to write a good essay, he or she may have to read a lot of good materials for preparation of expressions, wording choices as well as ideas.

4 Related Work

To the best of our knowledge, there exist few studies on automatically challenging the Chinese

[3] The standard deviation reflects the variance of different evaluators on each output, therefore also reflects agreements.

曲折(intricate) - 坎坷(bumpy) - 漫长(long-term) - 艰辛(hardship) - 历程(progress)
勇敢(brave) - 毅力(determination) - 勇气(courage) - 坚强(adamancy) - 坚韧(tenacity) - 自信(confident) - 信心(confidence)
意志(willpower) - 斗志(fighting will) - 顽强(indomitable) - 艰难(tough) - 困苦(tribulation)
可怕(dreadful) - 悲剧(tragedy) - 灾难(disaster) - 灾害(calamity) - 严峻(rigorous) - 因素(factors)
特殊(special) - 困难(difficulty) - 困境(straits) - 低谷(trough)
人生(life) - 理想(cause) - 精神(spirit) - 理念(ethic) - 思维(thinking) -观念(concept)
态度(attitude) - 理性(rational) - 冷静(calm) - 理智(wise)

(a)

曲折(intricate) - 坎坷(bumpy) - 漫长(long-term) - 艰辛(hardship) - 历程(progress)
可怕(dreadful) - 悲剧(tragedy) - 灾难(disaster) - 灾害(calamity) - 严峻(rigorous) - 因素(factors)
特殊(extraordinary) - 困难(difficulty) - 困境(straits) - 低谷(trough)
意志(willpower) - 斗志(fighting will) - 顽强(indomitable) - 艰难(tough) - 困苦(tribulation)
痛苦(suffering) - 悲伤(sorrow) - 伤痛(pain) - 痛楚(agony)
悲哀(grieve) - 难过(sad) - 失望(disappointed) - 绝望(despair) - 无助(helpless) - 孤独(loneliness)
茫然(at a loss) - 迷茫(confused) - 困惑(puzzled) - 尴尬(embarrassed)

(b)

Figure 1: Lexical chains formulated by (a) DivRank and (b) PageRank

college entrance tests. One recent work focuses on multiple choice questions in that context (Guo et al., 2017).

The approach of selecting sentence for constructing essays share similar methodological nature with extractive summarization, where classic graph-based ranking has been shown useful (Erkan and Radev, 2004). Diversified selection could further improve information coverage (Mei et al., 2010; Lin and Bilmes, 2011; Hong et al., 2014). Note that the goal of essay writing is different with summarization. The task in this study is to generate a rich but coherent article, and every student or system could write a very different essay, while the goal of summarization is to condense documents, in which case the output results should be similar in content, covering the same important facts.

The closest study with the main theme of this paper is perhaps the recent work by Qin et al. (2015) on essay generation, which directly utilizes words as subtopic representations rather than our proposed usage of diversified lexical chains.

5 Conclusion and Future Work

In this paper, we study the challenging task of essay construction for Chinese college entrance exams, propose a framework based on diversified lexical chains and show its effectiveness.

Our framework is simple in nature and is by no means perfect. For example, structural coherence is not explicitly modeled in our method since lexical chains could only capture topical coherence. We leave more elaborated strategies for content planning as future work. Also, we would like to extend the framework for more difficult titles or topics by exploring proper vectorial representations, and to collect manual data for supervised learning. Methods beyond sentence extraction should also be explored to utilize more elaborative syntactic and discursive structures.

Acknowledgments

This work was supported by 863 Program of China (2015AA015403), NSFC (61331011), and Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology). We thank the anonymous reviewers for helpful comments. Xiaojun Wan is the corresponding author.

References

Regina Barzilay and Michael Elhadad. 1999. Using lexical chains for text summarization. *Advances in automatic text summarization*, pages 111–121.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391.

Günes Erkan and Dragomir R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artifitial Intelligence Research (JAIR)*, 22:457–479.

David Graff and Ke Chen. 2005. Chinese gigaword. *LDC Catalog No.: LDC2003T09, ISBN*, 1:58563–58230.

Shangmin Guo, Xiangrong Zeng, Shizhu He, Kang Liu, and Jun Zhao. 2017. Which is the effective way for gaokao: Information retrieval or neural networks? In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 111–120, Valencia, Spain. Association for Computational Linguistics.

Kai Hong, John Conroy, Benoit Favre, Alex Kulesza, Hui Lin, and Ani Nenkova. 2014. A repository of state of the art and competitive baseline summaries for generic news summarization. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1608–1616, Reykjavik, Iceland.

Hui Lin and Jeff Bilmes. 2011. A class of submodular functions for document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 510–520, Portland, Oregon, USA. Association for Computational Linguistics.

Qiaozhu Mei, Jian Guo, and Dragomir Radev. 2010. Divrank: the interplay of prestige and diversity in information networks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1009–1018. Acm.

Jane Morris and Graeme Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational linguistics*, 17(1):21–48.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Bing Qin, Duyu Tang, Xinwei Geng, Dandan Ning, Jiahao Liu, and Ting Liu. 2015. A planning based framework for essay generation. *arXiv preprint arXiv:1512.05919*.

Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky, and Christopher Manning. 2005. A conditional random field word segmenter for sighan bakeoff 2005. In *Proceedings of the fourth SIGHAN workshop on Chinese language Processing*, volume 171. Citeseer.