# Fostering Students' Idea Generation through Corpus-based Recommendation in Online Writing Environment

**Calvin C. Y. LIAO[a*], Hercy N. H. CHENG[b] & Wan-Chen CHANG[c]**

[a] *College of Nursing, National Taipei University of Nursing and Health Sciences, Taiwan*
[b] *National Engineering Research Center for E-Learning, Central China Normal University, China*
[c] *Department of Human Development and Family Studies, National Taiwan Normal University, Taiwan*
*CalvinCYLiao@gmail.com

**Abstract:** Idea generation plays an important role in the writing process. The use of natural language techniques to mine semantic relationships between vocabularies or sentences in order to provide students with more suitable contents for writing ideas. Therefore, this study developed a corpus-based Chinese writing recommendation system. This study collects, analyzes, processes, and constructs a primary school corpus with 580,000 written texts, and develops a vocabulary and sentence recommendation mechanism for writing. 37 fourth-year students are invited to participate in the evaluation. The study found that students provided 60% of the candidate vocabularies or sentences they provided, but 40% of the candidate vocabularies or sentences were not applicable. Furthermore, students had more than 60% positive attitude towards the system.

**Keywords:** idea generation, writing recommendation, word vector, sentence similarity computation

## 1. Introduction

Writing is one of the essential components of language learning. The creative and comprehensive characteristics of the writing process determine the overall quality of students' overall development. Specifically speaking, writing can promote the improvement and development of students' psychological quality, cultural quality, and thoughts and feelings (Flower, & Hayes, 1981). However, in fact, students often need to write under the similar life situations and specific requirements, but it is challenging to write articles that are consistent with the meaning of the questions, the structure is reasonable, and the sentences are fluent. Therefore, students often face vocabulary choices during the writing process, the expressions are unclear, and the long story is not known (Hayes, & Flower, 1980). This may be because students are not sure what they want to express. The cognitive writing model (Fitzgerald, & Teasley, 1986) points out that idea generation is the most basic element in the writing process. Furthermore, the process of generating writing ideas is the exploration of students' long-term memory based on the writing theme. Therefore, it is not only the creation of writing ideas but also how to help students present these ideas in the form of words.

With the development and popularization of network technology, the writing style has gradually changed from the traditional paper-and-pencil mode to the digital and mobile mode. The natural language processing technology in the field of artificial intelligence plays a vital role in language engineering and cognitive science. Therefore, this study attempts to design a corpus-based primary school vocabulary and sentence recommendation mechanism to assist and support student writing. Specifically, this study intends to develop a personalized writing recommendation mechanism. Through the analysis of the massive writing text corpus, the vocabulary and sentence correlation database are established. When students write, they are recommended to have semantic relations with the current keywords. Vocabulary or excellent sentences with similar themes to the current sentence make the writing process more organically accessible to different but related ideas, and students apply it to their writing. Furthermore, this study focuses on the mining of semantic

relations between vocabulary, the similarity judgment based on the subject between sentences, and the construction of text personalized recommendation system (Poesio, 2011). Besides, how to use the developed writing vocabulary and sentence recommendation mechanism to guide, or in what form to present and guide, is still a challenge that needs to be solved urgently.

## 2. System Development

### 2.1 Article Collection and Data Verification

The establishment of the corpus is mainly based on Chinese essays in primary and secondary schools. There are 11 websites which offer a wealth of writing corpus resources. This study has written a crawler program in Python to crawl the content of these websites and to parse and extract corpus resources such as reading articles and writing texts. Subsequent classification of the acquired corpus, for example, a total of 5 writing styles, nine groups of grades, etc., a total of 620,000 articles.

### 2.2 Chinese Corpus Building

This study compares the NLPIR developed by Professor Zhang Huaping of the Chinese Academy of Sciences, the open source tool HanLP, the well-known Stanford NLP in the field of English natural language processing, and the Jieba word segmentation based on the Python language. Finally, the NLPIR word segmentation tool is used to process the corpus and build a vocabulary library. After processing, the final data of 580,000 articles are used to construct the corpus. Continued, word-level corpus processing and mining include the selection and use of Chinese word segmentation tools, vocabulary distribution information statistics and analysis, and the training of word vector models. Sentence-level corpus processing and mining include two parts: the construction of sentence libraries and the topic modeling of sentences.

### 2.3 Design on Corpus-based Writing Recommendation

The system provides students with online writing, and the recommended functions of related vocabulary and sentences. The process is as follows: 1) *select the target*: the original vocabulary/sentence selected by the student mouse to make synonyms and related vocabularies (the recommendation of similar sentences, considering the system feedback time and the actual application effect, the system recommends ten associated vocabularies/sentences according to each vocabulary/sentence selected by the students; 2) *calculating the similarity*: the method according to vocabularies/sentences. The similarity (computed using the cosine distance) is used to obtain the first 20 candidate vocabularies/sentences; 3) *refer to the previous data and adjust*: the system compares the similarity calculation results of the candidate vocabularies/sentences, the number of praises to the students The historical writing record are combined to calculate the comprehensive score for the candidate vocabularies/sentences; 4) presenting the recommended content: the system performs the descending order to select the top 10 candidate vocabularies/sentences as the recommended option for the students; 5) judge the recommended content: Students will click "Like" or "X" (exclude) on the recommendation results to adjust the accuracy of the vocabulary/sentence) recommendation.

## 3. Preliminary Evaluation

### 3.1 Research Design and Procedure

The participants were 37 10-year-old fourth-grade students from an elementary school in Hubei, China. All students participated in the study over the period of one semester in Chinese literacy course. One goal of the Chinese literacy course is expected to help students learn how to write. Topic Chinese New Year Life was provided in Chinese. The goal was to understand the influence of OWE

with idea recommendation on students' writing behavior and attitude. The participants wrote same topic articles within their writing environment during six 60-min. When the students finished the experimental activity, attitude were administered to collect students' perceptions related to system usage (Davis, Bagozzi, & Warshaw, 1989).

## 3.2 Findings

*Behaviors:* This study analyzes the usage record of the vocabulary recommendation function. Each student uses 6.19 recommendations, clicks "Like" 4.27 times, and clicks "X" 2.3 times. For the vocabulary such as synonyms and related words provided by the system, there is about 68.98% recognition, but 37.16% of the recommendations are still not applicable. Analyze the usage record of the sentence recommendation function. Each student uses 3.24 recommendations, clicks "Like" 1.81 times, and clicks "X" 1.68 times. There are about 55.86% acceptance of the candidate sentences provided by the system, but there are still 51.85% of the sentence recommendations not applicable. This means that the relevance of the sentence recommendation results needs to be improved.

*Attitudes:* The results of writing attitudes indicate that usefulness (nearly 60% consent) and ease of use (almost 60% consent), the writing recommendation function not only has the function of improving students' writing skills, improving students' writing efficiency, assisting students' writing, but also convenient. Easy-to-use features; from the perspective of usage attitude (nearly 60% consent), willingness to use (nearly 70% consent), and student growth needs (almost 70% consent), writing with this assisted writing system is delightful. Therefore, students are pleased to try and use.

## Acknowledgements

## References

Chou, C.Y. and Lu, L. (2014). Exploring the Attitude Differentiation on e-Learning Systems Based on TAM: the Strength of Growth Need as A Moderator, *Journal of Information Management*, 21(1), 83-106.

Burgess, C., Livesay, K., & Lund, K. (1998). Explorations in context space: words, sentences, discourse. *Discourse Processes, 25*(2-3), 211-257.

Davis, F.D., Bagozzi, R.P. and Warshaw, P.R. (1989). User acceptance of computer technology: a comparison of two theoretical models, *Management Science*, Vol. 35, No. 8, pp. 982-1003.

Fitzgerald, J., & Teasley, A. B. (1986). Effects of instruction in narrative structure on children's writing. *Journal of Educational Psychology, 78*(6), 40.

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition & Communication, 32*(4), 365-387.

Hayes, J. R., & Flower, L. S. (1980). *Identifying the organization of writing processes. Cognitive processes in writing*.

Poesio, M. (2011). Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research, 11*(1), 95--130.