Ogata, H. et al. (Eds.) (2015). Proceedings of the 23rd International Conference on Computers in Education. China: Asia-Pacific Society for Computers in Education

Enhanced Bilingual Text Analysis for BYOD with Hierarchical Visualization

Ping LI*, Siu Cheung KONG*, Tak-Lam WONG & Chengwei GUO*
Department of Mathematics and Information Technology
The Hong Kong Institute of Education, Hong Kong
*{pli, sckong, guoc}@ied.edu.hk

Abstract: Nowadays, bring your own device is making significant inroads and becoming one of the hottest research trends in higher education. To enhance teaching and learning efficiency and promote mobile learning quality in Hong Kong, this paper proposes an enhanced bilingual text analysis system with hierarchical visualization, in which a text corpora of discussions from students and a set of concepts (topics) with keywords designed by teachers in a Moodle course will be automatically collected through a mobile learning platform, before being processed for further text analysis. A hierarchical keywords visualization after discussion text analysis in a distinct period will be utilized and the statistics of the students' access to keywords will be displayed in a hierarchical structure manner to reflect the learning effects. Our system will be useful for teachers to check students' learning progress and further optimize their teaching strategies even a massive amount of text data is generated in discussions. The distinguishing feature of our system is that it supports bilingual analysis for mixed language text of English and Chinese, addressing a key issue in Hong Kong that the two languages are commonly used in a mixed manner in daily communication.

Keywords: Bilingual text analysis, hierarchical structure, keywords visualization

1. Introduction

The teaching strategy of Bring Your Own Device (BYOD) is recently very popular and attracting more and more teachers to encourage their students to apply their own personal devices like laptops, tablets and smart phones as helpful assistant tools for both in-class and off-class mobile learning. However, since the availability of mobile tools is highly increased and online access to discussion forums via BYOD becomes more convenient, an explosion of data size has turned out to be a new serious problem for teaching and learning in mobile education settings. Students tend to make new posts on online forums whenever they like, and it leaves teachers with a large amount of discussion data for understanding the learning progress of students, which will be a tedious, time consuming or even impossible task for teachers concerning the data size. Then, how about leaving the Social Network Platforms (SNPs) and Learning Management Systems (LMSs), and going back to traditional pedagogical practices while other teachers and students are getting along well with new technologies? This is of course not a good answer to the problem, since it is definitely behind the current trend in the modern society, especially in an era of big data in which almost everything is online and every discussion is massive. In a special education environment of Hong Kong, both Chinese and English serve as officially languages in daily communication. The students will make posts in online discussion forums either in Chinese or in English or even in a more complicated combined bilingual manner just when they consider it's convenient for them to express their personal ideas in communication. Such features of discussion contents add great challenges in the task of analyzing and understanding the discussions by machine automation in addition to the current massive data size environments.

With the aim of enhancing teaching and learning efficiency/quality on e-learning platforms and LMSs in Hong Kong, as well as promoting mobile learning using BYOD, this paper proposes an enhanced bilingual text analysis with hierarchical keywords visualization, which supports text processing with mixed language discussion contents as input. Teachers will design several key concepts (topics) for students to discuss before teaching. Under each concept (topic), several keywords will also be designed to check whether the students' discussions cover each concept (topic) well. The topics and keywords designed will also be available in both English and Chinese. We provide a Latent Dirichlet Allocation (LDA) based keywords recommendation in our system for giving suggestions to teachers to help them design the topics and keywords that they want the students to cover in the discussions. Both pre-course and post-course discussions will be conducted in our experiments. A hierarchical visualization for both the pre-course and post-course discussions showing clearly the structural relation of topics and keywords will be presented in our system. Students access to keywords will be clearly displayed showing the learning effects before and after the course study, which will be very useful for teachers to understand the learning outcome of students and further optimize the teaching strategies even massive data are generated via BYOD.

2. Related Work

Text Analytics This paper is made possible by the inspiration of previous work. Concerning the special environment of Hong Kong, where both Chinese and English serve as official languages in communication, bilingual text analysis, namely, the English keywords discovery and the Chinese keywords discovery is supported in our system. Generally, English keywords detection is much simpler due to the word meaning unitary, although segmentation may be needed for some English words, yet usually simple tasks compared with Chinese word segmentation (Fu & Luke, 2003; Gao, Li, & Huang, 2003; Peng, Feng, & McCallum, 2004; Yang, Lee, & Chen, 2009). Different methods have been proposed by researchers in the fields related to computational linguistics to solve the problems in Chinese word segmentation. A two-stage statistical word segmentation for Chinese based on word bigram and word formation models is proposed (Fu & Luke, 2003). A linear-chain conditional random field method is performed to analyze the Chinese word segmentation integrating domain knowledge in the form of multiple lexicons of characters and words (Peng, Feng, & McCallum, 2004), and an improved probabilistic new word detection method is provided. A Chinese word segmentation method which applies improved source-channel models is introduced through dividing Chinese words into four types: lexicon words, morphologically derived words, factoids and named entities (Gao, Li, & Huang, 2003). These methods provide us with inspiration for Chinese keywords extraction, while English keywords extraction is much simpler (Chitrakala & Manjula, 2007).

More advanced methods for Chinese text analytics are further presented allowing more dynamic and accurate analysis of Chinese sentences. As words in Chinese sentences are not naturally separated by delimiters, Chinese keywords detection in text for online discussion board is a great challenge (Xu, Gao, Toutanova, & Ney, 2008). Adaptive Chinese word analysis (Gao et al., 2004) is presented for different domains and standards, where domain-specific words are applied for word segmentation based on classic linear models. Factors like segmentation consistency and granularity of Chinese words are found considered as more essential in (Chang, Galley, & Manning, 2008) for machine translation. To enhance the segmentation effectiveness, a decoding method for discriminative joint Chinese word analytics and parsing is proposed (Qian & Liu, 2012). High dimensional characters like word-based features and enriched edge features are integrated for joint analytics modeling (Sun, Wang, & Li, 2012). A two-step stacked sub-word model for joint Chinese word analytics is presented considering both the efficiency and effectiveness (Sun, 2011). A pragmatic approach for Chinese word analytics is investigated based on the usage of the Chinese word in real computer applications, pragmatic mathematical framework, and different granularities of Chinese words (Gao, Li, Wu, & Huang, 2005). A character-based tagging problem is presented and nice segmentation is achieved for text analytics (Zhao, Huang, Li, & Lu, 2010). Further, a semi-supervised domain adaptation algorithm (Wang, Zong, & Su, 2012) is proposed for enhancing cross-domain corpora performance. Although English text analysis is well optimized and also there are many methods for processing with difficult Chinese text analysis, yet there are only few researches dealing with bilingual text analysis, namely, the text analytics for the mixed language input of English and Chinese. However, concerning the prevalence of this kind of language usage in Hong Kong, it is necessary and a must to solve this critical issue in our bilingual text analysis system.

Visualization Representation Different visualization styles will achieve different understanding purposes. As known to all, the same dataset represented and visualized in different manners will convey totally different knowledge of focuses to impact the creation of insight understanding (Moere, Tomitsch, Wimmer, Christoph, & Grechenig, 2012). One simple case is the collaboration diagrams (Abdurazik & Offutt, 2000) and the sequence diagrams (Grønmo, Runde, & Møller-Pedersen, 2013) in software engineering. Generally, they are fully equivalent interaction diagrams illustrating how objects interact with each other, all features of the sequence diagrams can be applied to the collaboration diagrams equivalently (Schach, 2011). However, when the organization of objects is the focus, a collaboration diagram will be used, and when the transmission of message is the focus, a sequence diagram will be applied. Elmqvist and Fekete (2010) introduced a model for constructing and visualizing with multi-scale representations of information visualization via hierarchical aggregation. Teoh and Ma (2002) proposed a technique for visualizing large hierarchies, which is a new ringed circular layout of nodes for making efficient use of limited display space. Since we also want to produce a hierarchical visualization for the representation of discussions keywords illustrating clearly the structural relation of topics and keywords in our system, where students' access to keywords will be clearly displayed showing out the learning effects for pre-course and post-course discussions, it is necessary and possible to apply the latest visualization techniques (Bostock, Ogievetsky, & Heer, 2011) to build a hierarchical visualization for aesthetical scientific visualization of discussion keywords. Hence, we will be able to know the knowledge coverage of students in discussion forums well.

3. Bilingual Analytics and Hierarchical Visualization Framework

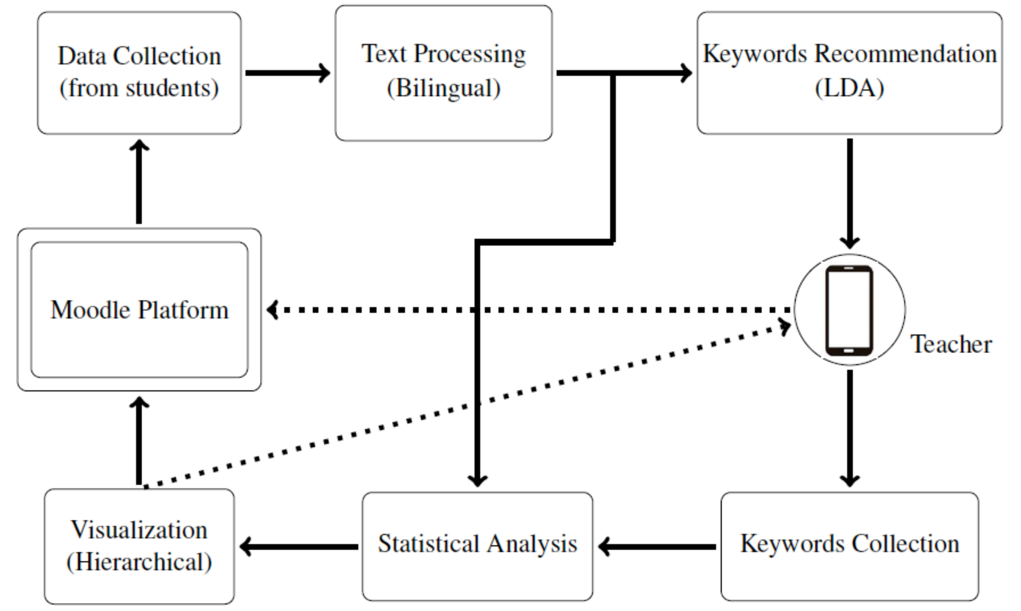
Data Collection (from students)
Text Processing (Bilingual)
Keywords Recommendation (LDA)
Moodle Platform
Teacher
Visualization (Hierarchical)
Statistical Analysis
Keywords Collection

Figure 1. The Framework of Our Bilingual Text Analysis System with Hierarchical Visualization.

We design a Moodle-based platform with mobile functions to collect students' discussions together with concepts (topics) and keywords designed by teachers. Furthermore, we apply text mining techniques to bilingually analyze the collected discussion data. The analysis results will be shown on our platforms through a hierarchical visualization approach. The general framework of our system is shown in Figure 1. In our system, teachers are asked to design key concepts (topics) containing several keywords for evaluating the qualities of their students' discussions and identifying the weakness of students' learning scopes. There might be some proper words in the discussions which are synonymous to the keywords designed by teachers, but these words will not be counted as they do not exist in the keywords library. This issue will lead to biased criteria in the analysis, hence our keywords recommendation function will be useful for teachers to refine and enrich their keywords. For such purpose, we extract potentially useful topics and related keywords from students' discussions by using Latent Dirichlet Allocation (LDA). LDA is a generative probabilistic model for collections of discrete data such as text documents, and was first presented as a graphical model for topic discovery (Blei, Ng, & Jordan, 2003). It is a way of automatically discovering topics and keywords from documents. Since the input discussions are not suitable to conduct a cluster analysis, we first pre-process the input discussions: apply Part-of-Speech Tagging (for both English and Chinese) on the text, and extract words with Part-of-Speech - noun, verb, and adjective. Furthermore, we apply LDA algorithm on extracted text to find keywords and constitute an important keywords recommendation for teachers.

4. Bilingual Text Processing

As we discussed before, the environment of Hong Kong is special as it allows bilingual communication in teaching and learning. As Chinese and English are daily used in both oral and written situations, it is therefore common to find that in some discussions, students will mix Chinese words or phrases into an English sentence or English words or phrases into a Chinese sentence. To solve this problem, we first implement ENGLISH TEXT PROCESSING (Klein & Manning, 2003; Toutanova, Klein, Manning, & Singer, 2003) and CHINESE TEXT PROCESSING (Chang, Galley, & Manning, 2008) individually from the input language-mixed text, then apply a combinatorial approach to complete the bilingual text analysis. The input discussion text that we deal with in our system are mixtures of English and Chinese words. Almost every sentence in the text has the form consisting of a sequence of English words mixed up with Chinese phrases, or a sequence of Chinese words mixed up with English phrases. For each bilingual sentence, we first check its domain language (i.e. the language that most of words are in), then apply level-1 word segmentation for domain language on the sentence, and finally apply level-2 word segmentation afterwards on the sub-sequences of consecutive words in the sentence that are in another language. The outline of our approach for BILINGUAL TEXT PROCESSING is shown in Figure 2.

```
BILLINGUAL TEXT PROCESSING
Input:   A sentence $s$ of bilingual words.
Output:  A sequence of groups of words.
if Domain language of $s$ is English then
    Partition $s$ into groups $w = e_1 e_2 \cdots e_k$ by ENGLISH TEXT PROCESSING.
    for each maximal consecutive Chinese words $e_p e_{p+1} \cdots e_q$ in $w$ do
      Make $e_p e_{p+1} \cdots e_q$ into $c'_1 c'_2 \cdots c'_l$ by CHINESE TEXT PROCESSING.
      $w \leftarrow e_1 \cdots e_{p-1} c'_1 \cdots c'_l e_{q+1} \cdots e_k$.
    Output $w$.
elseif Domain language of $s$ is Chinse then
    Partition $s$ into groups $w = c_1 c_2 \cdots c_m$ by CHINESE TEXT PROCESSING.
    for each maximal consecutive English words $c_s c_{s+1} \cdots c_t$ in $w$ do
      Make $c_s c_{s+1} \cdots c_t$ into $e'_1 e'_2 \cdots e'_n$ by ENGLISH TEXT PROCESSING.
      $w \leftarrow c_1 \cdots c_{s-1} e'_1 \cdots e'_n c_{t+1} \cdots c_m$.
    Output $w$.
endif
```

Figure 2. The Outline of Our Algorithm for BILINGUAL TEXT PROCESSING.

5. Experimental Results

In the course "E-Learning in Primary Schools" offered by HKIEd in 2014/15 Semester II, our bilingual text analytics system has been utilized in order to enhance the teaching and learning effects, also to test and evaluate the system. There are totally 40 students in the course. Students are required to input their understandings of the definition of e-learning twice as discussions in the Moodle forum: at the start of the course (pre-course), and after the training of the course (post-course). Both English and Chinese are allowed in the discussions. The teacher determines the keywords after the first discussion period. These data are then collected through our Moodle-based platform. Our text analysis results are shown using hierarchical keywords visualization in Figure 3, the analysis for the pre-course discussion is shown in the left circle in Figure 3, while the result of the post-course discussion is shown in the right circle in Figure 3. Figure 3 hence shows the clear changes in the students' understanding between the pre-course and post-course discussions. Through our hierarchical visualization, we can clearly see that after the course learning, students have better and more comprehensive understandings of the concepts in e-learning. Furthermore, comparing the pre-course (left circle) and post-course (right circle) discussions, we can find that, in the early stage of the course, students almost know nothing about several concepts (topics) such as “Principles/Models/Theories”, “Digital ways of communication”, and such situation is improved after the course training. However, there are still some topics of keywords are concerned by few students such as “Assessment”, “Data collection and analysis”, etc. Based on such observation, teacher provides complementary materials for students’ learning. This also corresponds to the original aim and motivation for us to design this teaching and learning enhancement system.

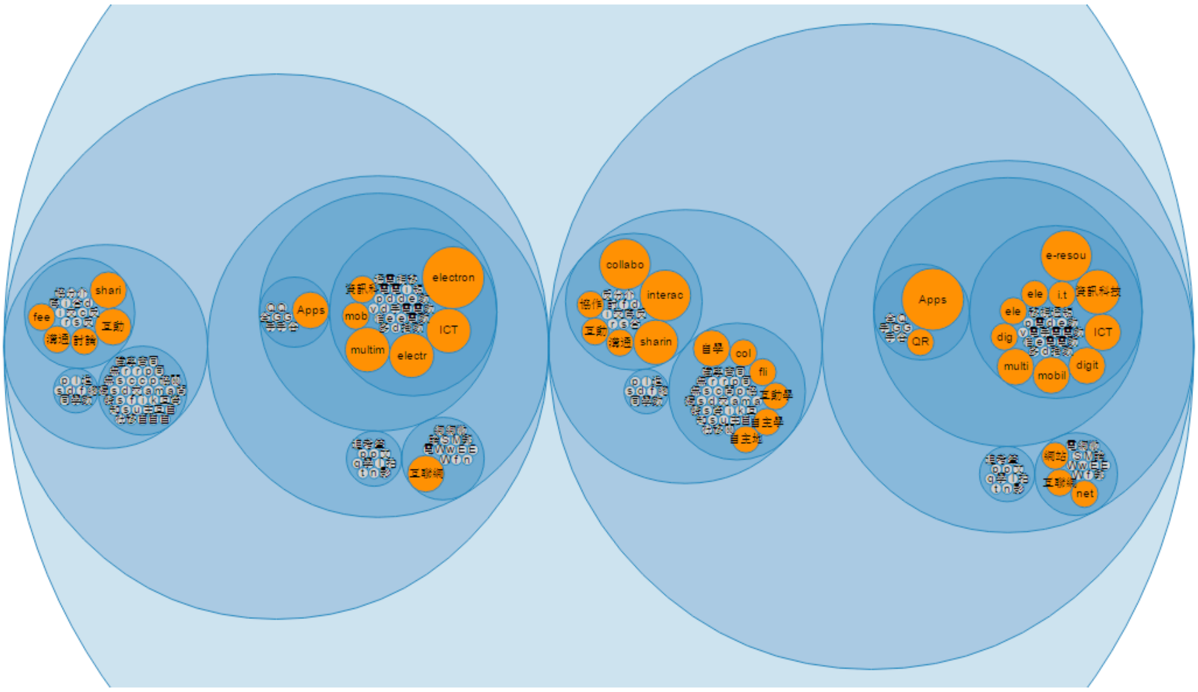

Figure 3. E-Learning Class: Pre-Course (Left Circle) and Post-Course (Right Circle) Discussions.

6. Summary

This study proposed an enhanced bilingual text analytics system with hierarchical keywords visualization, which supports text processing with mixed language input of English and Chinese. As a promotion of BYOD strategy, a mobile learning based Moodle platform was utilized for collecting data from students and teachers. The main contribution of this study is to apply the bilingual text analysis techniques in order to present a hierarchical visualized analysis of students’ learning discussion to teachers for further enhancing the teaching and learning quality. The system has been applied to real course teaching in HKIEd, and satisfactory results and feedbacks are received. We believe that this study has provided a meaningful inspiration and could be a seminal contribution to the further researches on bilingual analysis in teaching and learning. We will further work on a more intelligent spelling correction mechanism for bilingual text analytics to extend our system in order to improve the accuracy of the text analysis results. Moreover, this text analysis system can be further developed by enhancing the intelligence of the statistical analysis model. Based on a specific data set of students’ discussions and teachers’ criteria for training set, latest machine learning techniques could be incorporated to improve the quality of scoring mechanism in the statistical analysis phase.

Acknowledgements

The authors would like to thank Prof. Ulrich Hoppe for his helpful suggestions to the work, and acknowledge the support of the HKSAR UGC Funding (ref. HKIEd7/T&L/12-15).

References

Abdurazik, A., & Offutt, J. (2000). Using UML collaboration diagrams for static checking and test generation. *Proceedings of the 3rd International Conference on the Unified Modeling Language: Advancing the Standard*, 383-395.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993-1022.

Bostock, M., Ogievetsky, V., & Heer, J. (2011). D3: data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12), 2301-2309.

Chang, P.-C., Galley, M., & Manning, C. D., (2008). Optimizing Chinese word segmentation for machine translation performance. *Proceedings of the Workshop on Statistical Machine Translation*, 224-232.

Chitrakala, S., & Manjula, D. (2007). Distributed multi-lingual content based text mining DML-CBTM. *Proceedings of the 3rd conference on IASTED International Conference: Advances in Computer Science and Technology*, 500-505.

Elmqvist, N., & Fekete, J.-D. (2010). Hierarchical aggregation for information visualization: overview, techniques, and design guidelines. *IEEE Transactions on Visualization and Computer Graphics*, 16(3), 439-454.

Fu, G., & Luke, K. K. (2003). A two-stage statistical word segmentation system for Chinese. *Proceedings of the SIGHAN Workshop on Chinese Language Processing*, 17, 156-159.

Gao, J., Li, M., & Huang, C.-N. (2003). Improved source-channel models for Chinese word segmentation. *Proceedings of the Annual Meeting on Association for Computational Linguistics*, 1, 272-279.

Gao, J., Li, M., Wu, A., & Huang, C.-N. (2005). Chinese word segmentation and named entity recognition: a pragmatic approach. *Computational Linguistics*, 31(4), 531-574.

Gao, J., Wu, A., Li, M., Huang, C.-N., Li, H., Xia, X., & Qin, H. (2004). Adaptive Chinese word segmentation. *Proceedings of the Annual Meeting on Association for Computational Linguistics*, 462:1-462:8.

Grønmo, R., Runde, R. K., & Møller-Pedersen, B. (2013). Confluence of aspects for sequence diagrams. *Software and Systems Modeling*, 12(4), 789-824.

Klein D., & Manning, C. D. (2003). Accurate unlexicalized parsing. *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, 1, 423-430.

Moere, A. V., Tomitsch, M., Wimmer, C., Christoph, B., & Grechenig, T. (2012). Evaluating the effect of style in information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 18(12), 2739-2748.

Peng, F., Feng, F., & McCallum, A. (2004). Chinese segmentation and new word detection using conditional random fields. *Proceedings of the International Conference on Computational Linguistics*, 562:1-562:7.

Qian, X., & Liu, Y. (2012). Joint Chinese word segmentation, POS tagging and parsing. *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 501-511.

Schach, S. R. (2011). *Object-oriented and classical software engineering* (8th ed.). New York: McGraw-Hill.

Sun, W. (2011). A stacked sub-word model for joint Chinese word segmentation and part-of-speech tagging. *Proceedings of the Annual Meeting of The Association for Computational Linguistics: Human Language Technologies*, 1, 1385-1394.

Sun, X., Wang, H., & Li, W. (2012). Fast online training with frequency-adaptive learning rates for Chinese word segmentation and new word detection. *Proceedings of the Annual Meeting of the Association for Computational Linguistics: Long Papers*, 1, 253-262.

Teoh, S. T., & Ma, K.-L. (2002). RINGS: a technique for visualizing large hierarchies. *Proceedings of the Revised Papers from the 10th International Symposium on Graph Drawing*, 268-275.

Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, 1, 252-259.

Wang, K., Zong, C., & Su, K.-Y. (2012). Integrating generative and discriminative character-based models for Chinese word segmentation. *ACM Transactions on Asian Language Information Processing*, 11(2), 7:1-7:41.

Xu, J., Gao, J., Toutanova, K., & Ney, H. (2008). Bayesian semi-supervised Chinese word segmentation for statistical machine translation. *Proceedings of the International Conference on Computational Linguistics*, 1, 1017-1024.

Yang, H.-C., Lee, C.-H., & Chen, D.-W. (2009). A method for multilingual text mining and retrieval using growing hierarchical self-organizing maps. *Journal of Information Science*, 35(1), 3-23.

Zhao, H., Huang, C.-N., Li, M., & Lu, B.-L. (2010). A unified character-based tagging framework for Chinese word segmentation. *ACM Transactions on Asian Language Information Processing*, 9(2), 5:1-5:32.