

REDUCING DOCUMENT TEXT TO HIGHLIGHT THE ESSENCE OF THE MAJOR POINTS OF THE ORIGINAL DOCUMENT

S.RENUKA¹, K.CHANDHARA MOULI²

¹Student, NRI INSTITUTE OF TECHNOLOGY, A.P., India.

²Associate Professor & HOD, Dept. Of Department of Master of Computer Applications, NRI
INSTITUTE OF TECHNOLOGY, A.P., India.

ABSTRACT: Text highlighting chooses those sentences verbatim from the text that provide an overview of the entire passage. Automatic Text Summarization is the process of reducing document text to highlight the essence of it and retain only the major points of the original document. There are two different approaches to create these summaries – extractive text summarization and abstractive text summarization. Extractive summaries are created by choosing sentences from the text that are necessary to capture the meaning of it while ignoring those sentences that can be removed without losing the meaning of the given text. It does not generate any new sentences. This paper addresses the problem of predicting new highlights of partly highlighted electronic learning documents. With the goal of enriching teaching content with additional features, text classification techniques are exploited to automatically analyze portions of documents enriched with manual highlights made by users with different levels of knowledge and to generate ad hoc prediction models. Then, the generated models are applied to the remaining content to suggest highlights. To improve the quality of the learning experience, learners may explore highlights generated by models tailored to different levels of knowledge. We tested the prediction system on real and benchmark documents highlighted by domain experts and we compared the performance of various classifiers in generating highlights.

INTRODUCTION

Electronic textual documents are among the most popular teaching content accessible through e-learning platforms. Teachers or learners with different levels of knowledge can access the platform and highlight portions of textual content which are deemed as particularly relevant. The highlighted documents can be shared with the learning community in support of oral lessons or individual learning. However, highlights are often incomplete or unsuitable for learners with different levels of knowledge. Additionally, extractive summarization techniques have shown better results

than most abstractive summarization techniques since this technique does not modify the intent of the sentence, especially when the usage of a particular figure of speech is not common across languages or the way a specific language is spoken in different regions. Most abstractive systems use extractive techniques as well to improve the accuracy of the output. In this paper we address the issue of automatically generating document highlights. Highlights are graphical signs that are usually exploited to mark part of the textual content. For example, the most significant parts of the text can be underlined, colored, or circled. The importance of text highlights in learning activities has been confirmed by previous studies on educational psychology (e.g. [3]) and visual document analysis (e.g. [4]). The highlighted documents can be easily shared between teachers and learners through e-learning platforms [2]. However, the manual generation of text highlights is time-consuming, i.e., it cannot be applied to very large document collections without a significant human effort, and prone to errors for learners who have limited knowledge on the document subject. Automating the process of text highlighting requires generating advanced analytical models able to (i) capture the underlying correlations between textual contents and (ii) scale towards large document collections. The contribution of this paper is twofold: (1) It proposes to use text classification techniques to automate the process of highlighting learning documents. (2) It considers

the proficiency level of the highlighting users to drive the generation of new highlights.

LITERATURE SURVEY

1. Nallapatti et. Al [1].

Text Highlighting involves selecting certain important sentences from a document that help provide a good overview of the entire document. In this paper we explore five different machine learning models built to perform text highlighting.

2. J. L. Moore, C. Dickson-Deane, and K. Galyen, "E-learning, online learning, and distance

learning environments: It is not uncommon that researchers face difficulties when performing meaningful cross-study comparisons for research. Research associated with the distance learning realm can be even more difficult to use as there are different environments with a variety of characteristics. We implemented a mixed-method analysis of research articles to find out how they define the learning environment. In addition, we surveyed 43 persons and discovered that there was inconsistent use of terminology for different types of delivery modes. The results reveal that there are different expectations and perceptions of learning environment labels: distance learning, e-Learning, and online learning.

3. S. Saraswathi, M. Hemamalini, S. Janani, and

V. Priyadharshini: To improve the nature of the learning background, students may investigate features produced by models custom-made to various dimensions of information. We tried the

expectation framework on genuine and benchmark reports featured by space specialists and we looked at the execution of different classifiers in producing features. The accomplished outcomes exhibited the high precision of the forecasts and the appropriateness of the proposed way to deal with genuine instructing reports.

PROPOSED SYSTEM

We address the issue of automatically generating document highlights. Highlights are graphical signs that are usually exploited to mark part of the textual content. For example, the most significant parts of the text can be underlined, colored, or circled. The importance of text highlights in learning activities has been confirmed by previous studies on educational psychology and visual document analysis. The highlighted documents can be easily shared between teachers and learners through e-learning platforms. However, the manual generation of text highlights is time-consuming, i.e., it cannot be applied to very large document collections without a significant human effort and prone to errors for learners who have limited knowledge on the document subject. Automating the process of text highlighting requires generating advanced analytical models able to (i) capture the underlying correlations between textual contents and (ii) scale towards large document collections.

The manually highlighted documents are first collected into a training dataset. Some established text processing steps are then applied to

prepare the raw data to the next classification process. Classification entails learning a model from the subset of document sentences that have been manually highlighted by human experts. The model is exploited to analyze new sentences of the collection and decide whether they are worth being highlighted or not based on their content and, possibly, based on the level of knowledge of the highlighting user. Finally, learners are provided with highlights corresponding to different levels of knowledge.

IMPLEMENTATION

1. Data Representation

For each sentence of the training and test document collections we consider the following attributes: (i) the textual content, (ii) the presence of highlights, and (iii) the level of knowledge of the user who highlighted the sentence (if any). The training data consists of a set of records.

2. Text Preparation

To predict highlights from learning documents, the HIGHLIGHTER system considers the following features: (i) the occurrences of single terms (unigrams) in the sentence text, (ii) the occurrence of sequences of terms (n-grams), and (iii) the level of knowledge of the user who highlighted the sentence (if available). To properly handle textual features during sentence classification, few basic preparation steps are applied. First, non-

textual content occurring in the text is automatically filtered out before running the learning process. Then, two established text processing steps are applied: (i) stemming and (ii) stopword elimination.

3. Feature Selection

To predict the class value of the test records, features in the training dataset may have different importance. Some of them are strongly correlated with the class and, thus, their presence is crucial to perform accurate predictions. Others are uncorrelated with the class. Hence, their presence could be harmful, in terms of both accuracy and efficiency of the classification process.

4. Text Classification

Classification is a two-step process which entails: (i) Learning a model from the training dataset, called classifier, which considers the most significant correlations between the class and the other data features, and (ii) assigning a class value to each record in the test dataset, based on the previously generated model. To investigate the use of text classification algorithms in highlight prediction, we learn multiple benchmark classifiers relying on different techniques.

5. Per-Level Document Highlighting

If in the training dataset there is no information about the level of knowledge of the users, one single classification model is

generated and used to predict new highlights. Otherwise, the knowledge level of the highlighting users is considered because it is deemed as relevant to perform accurate highlight predictions.

CONCLUSION

This paper proposes highlighter, a new approach to automatically generating highlights of learning documents. It generates classification models tailored to different levels of knowledge from a set of highlighted documents to predict new highlights, which are provided to learners to improve the quality of their learning experience. A performance comparison between various classifiers on benchmark data and an analysis of the usability of the proposed approach on real document collections have been performed. In the current version of the system, highlights are not personalized. Specifically, the same highlights are deemed as appropriate for all the users having the same level of knowledge.

Future work

We aim at tailoring the automatically generated highlights to specific users. Therefore, we would like to generate not only unified and per-level models, but also user-centric models. Furthermore, we currently ignore the presence of textual annotations, which could enrich the document content with additional notes or rephrases. We plan to analyze such automatically generated

content to gain insights into the level of knowledge of learners.

REFERENCES

- [1] J. L. Moore, C. Dickson-Deane, and K. Galyen, "E-learning, online learning, and distance learning environments: Are they the same?" *The Internet and Higher Education*, vol. 14, no. 2, pp. 129 – 135, 2011.
- [2] F. Grunewald and C. Meinel, "Implementation and evaluation of digital e-lecture annotation in learning groups to foster active learning," *TLT*, vol. 8, no. 3, pp. 286–298, 2015.
- [3] S. Elliott, *Educational Psychology: Effective Teaching, Effective Learning*. McGraw-Hill, 2000.
- [4] A. B. Alencar, M. C. F. de Oliveira, and F. V. Paulovich, "Seeing beyond reading: A survey on visual text analytics," *Wiley Int. Rev. Data Min. and Knowl. Disc.*, vol. 2, no. 6, pp. 476–492, Nov. 2012.
- [5] C. Aggarwal and C. Zhai, "A survey of text classification algorithms," in *Mining Text Data*, C. C. Aggarwal and C. Zhai, Eds. Springer US, 2012, pp. 163–222.
- [6] D. D. Lewis, "Naive (bayes) at forty: The independence assumption in information retrieval," in *ECML*, 1998, pp. 4–15.
- [7] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, Mar. 1986.
- [8] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.
- [9] W. W. Cohen, "Fast effective rule induction," in *Proceedings of the Twelfth International Conference on Machine Learning*. Morgan Kaufmann, 1995, pp. 115–123.
- [10] E. Baralis, S. Chiusano, and P. Garza, "A lazy approach to associative classification," *IEEE TKDE*, vol. 20, no. 2, pp. 156–171, 2008.
- [11] Document Understanding Conference, "HTL/NAACL workshop on text summarization," 2004.
- [12] Y. H. Lee, G. D. Chen, L. Y. Li, Nurkhamid, C. Y. Fan, and K. H. Chiang, "The effect of utilizing the learning skill of highlighting and constructing a map in a networked hyperlink condition on learning performance," in *2012 IEEE 12th International Conference on Advanced Learning Technologies*, July 2012, pp. 546–548.
- [13] D. Suthers and K. Verbert, "Learning analytics as a "middle space"," in *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, ser. LAK '13. New York, NY, USA: ACM, 2013