Using text summarization models to improve digital reading of scientific papers

Ludovica Mastrobattista[1], Motasem Alrahabi[2], Valentina Fedchenko[2], Oussama Jomaa[3], James Gawley[2], Johanna Cordova[2], Glenn Roe[2].

1: l.mastrobattista@usal.es

2: prenom.nom@sorbonne-universite.fr

3: prenom.nom@fabriquenumerique.fr

Abstract: This paper presents an evaluation and comparison of three state-of-the-art models for text summarization, and proposes a new digital reading interface designed for neophyte users to exploit these models, as well as automatic keyword extraction, with little or no programming experience.

Keywords: digital reading, text summarization, keyword extraction, qualitative evaluation.

The scale of peer-reviewed publications in many fields is rapidly becoming unmanageable. As a result, academic researchers need new solutions to stay abreast of developments in their fields and take advantage of the latest research. Automatic text summarization (El-Kassas et al., 2021) represents an important opportunity to leverage automated reading techniques at scale, providing an overview of the study object, method and results, and to orient readers in the text even before reading it (Overstreet, 2021). This paper presents an evaluation and comparison of three state-of-the-art models for text summarization, and proposes a web interface designed for neophyte users to exploit these models without resorting to scripting or using the command-line interface. This study is part of a larger research project aimed at exploring the potential of digital environments that incorporate computational tools in order to provide alternative approaches to reading and improve the reader's performance and digital practice in an academic context.

Using a corpus of open-access academic articles related to the fields of communication and education from the journal *Comunicar: Scientific Journal of Media Education*[1], we evaluated the abstract summarizations produced by three existing models: BART (Lewis et al., 2019), PEGASUS (Zhang et al., 2020) and T5 (Raffel et al., 2020). We determined the evaluation criteria according to what in our opinion is a proper text assessment, which consists of three intrinsic measurements: the quality of the content, syntactic and morphological validity, and the accuracy of vocabulary. Based on these considerations, we conducted our evaluations using five rating levels developed according to the Mean Opinion Score (MOS) scale (Streijl et al., 2016; Iskender et al., 2021). This rating scale is typically expressed as a number between 1 (poor) and 5 (excellent).

[1] https://doi.org/10.3916/comunicar

Our evaluation[2] of 21 articles shows that the T5 model best summarizes the content of each section of the documents, according to the IMRaD structure (Sollaci & Pereira, 2004). T5 generated new data, i.e. sentences, from the original text extracting essential information and rendering the general sense of the source text coherently. On the other hand, the BART and PEGASUS models reached quite similar scores, that is around 3.5 for each summarized part of the paper. This means that the extracted summaries from both models present a level of quality between "fair" and "good", according to the MOS scale. Regarding the accuracy of vocabulary style, we noted that the majority of the extracted summaries copy-and-pasted sentences from the original text.

In order to take advantage of the affordances of a digital reading platform, allowing users to orient themselves in a text before reading it, we thought it would be interesting to cross-reference the abstract with another source of knowledge, i.e., the associated keyword for each part of the paper. In this way, an overview of the terminology of the text at hand is delivered by the combination of two sources: the summary and related keywords. KeyBERT, a keyword extraction method using the BERT embeddings model (Devlin et al., 2018), was used for this purpose. A second evaluation on seven articles shows that this option may improve the contextualisation of the article and the reader's orientation on the topic in question by providing relevant information present in the text from dual sources, the summary and the related keywords. In some cases, the keywords extracted from each part of the article contain multiple text-related terms that can then enrich the summary content.

To make these models available for testing and use by researchers without a strong technical background, we have created a freely accessible online interface[3]. This tool allows users to upload an article in XML format and receive an automatic summary of each section of the article, with the related keywords (the interface also allows for the summarization of plain text but in this case article titles are not taken into account). Different parameters can be selected to modify the size of the summary.

Our main scientific contribution in this paper is thus the qualitative evaluation of three important models for automatic summarization; and the development of an easy-to-use interface allowing researchers to experiment with these models through summaries generated along with keywords for each section of the analyzed article. We aim to evaluate these models on other sources of scientific articles, and to propose new functionalities which can facilitate digital reading and research.

[2] https://github.com/obtic-scai/Summarization
[3] https://obtic.sorbonne-universite.fr/summary/

References

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K., (2018). "Bert: pre-training of deep bidirectional transformers for language understanding". *arXiv:1810.04805v2 [cs.CL]*. https://doi.org/10.48550/arXiv.1810.04805

El-Kassas, W. S., Salama, C. R., Rafea, A. A., & Mohamed, H. K. (2021). “Automatic text summarization: A comprehensive survey”. *Expert Systems with Applications, 165,* 113679. https://doi.org/10.1016/j.eswa.2020.113679

Iskender, N., Polzehl, T., & Möller, S. (2021). "Reliability of Human Evaluation for Text Summarization: Lessons Learned and Challenges Ahead". In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)* (pp. 86–96).

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019). "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension". *arXiv:1910.13461 [cs.CL]*. https://doi.org/10.48550/arXiv.1910.13461

Overstreet M. (2021). "Networked Reading: How Digital Reading Experts Use Their Tools". *College English, 83* (5), pp. 357-378.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". *Journal of Machine Learning Research*, *21*, pp. 1-67. https://doi.org/10.48550/arXiv.1910.10683

Sollaci, L. B. & Pereira, M. G. (2004). "The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey". *J Med Libr Assoc*, *92* (3), pp. 364-7. PMID: 15243643; PMCID: PMC442179.

Streijl, R. C., Winkler, S., & Hands, D. S. (2016). "Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives". *Multimedia Systems 22*, pp. 213–227. https://doi.org/10.1007/s00530-014-0446-1

Zhang, J., Zhao, Y., Saleh, M., & Liu, P. J. (2020). "PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization". In *International Conference on Machine Learning, PMLR* (pp. 11328-11339). https://doi.org/10.48550/arXiv.1912.08777