Automatic Summarisation of Electronic Theses and Dissertations for Increased Media Engagement

Higher Education Institutions (HEIs) regularly publish manuscripts of academic research which provide useful insights into social, economic and technological issues affecting society (Phiri & M'sendo, n.d.). In the Electronic Theses and Dissertations (ETDs) researchers often provide focus on impact-driven research that has the potential to inform policy direction.

However, due to the large size of these ETD manuscripts, important stakeholders, such as local mainstream media outlets, find it difficult to synthesise the content of these manuscripts.

A potential solution would be to take advantage of advances in Artificial Intelligence by using Natural Language Processing (NLP) techniques to summarise the ETDs. Automatic text summarisation as Allahyari defined it, is the task of producing a concise and fluent summary while preserving key information content and overall meaning. The text summarisation techniques can be employed to generate snippets of scholarly research output that is concise and easy to assimilate by non-technical persons. (Allahyari et al., 2017; Ingram et al., n.d.). Existing literature broadly categories automatic summarisation into two broad classes: abstractive summarisation and extractive summarisation.

This study seeks to understand the challenges with synthesising ETD and design and implementation of software tools to be used to automatically summarise ETDs with a focus on mining text data and designing tools that allow for the automated summarisation and modification of the text, using usable tools.

Specifically, the objectives of this study are:

1. To determine how frequently research findings are reported in mainstream media.
2. To investigate challenges with synthesising long documents (ETDs).
3. To implement summarisation models for summarising ETDs for public consumption.
4. To evaluate the summarisation models.

The proposed research methodology is as follows:

- Frequency of publishing research findings—Manual and automatic content analysis of existing media publications will be conducted to determine the frequency of publishing research findings.
- Challenges synthesising ETDs—Interviews will be conducted with purposively sampled journalists from randomly sampled media outlets.
- Design and Implementation of ETD summarisation models—Classic text summarisation techniques such as abstractive and extractive summarisation will be employed in order to build the ETD summarisation models. Additionally, publicly available ETDs from HEIs in Zambia will be used to construct a dataset to be used during the study.

- Evaluation of ETD summarisation models—Standard evaluation metrics for text summarisation, such as ROUGE (Johnson, n.d.); BLEU (Allahyari et al., 2017; Johnson, n.d.); BERTScore (Bhandari et al., 2020; Zhang et al., n.d.) and METEOR (Ermakova et al., 2019; Johnson, n.d.), will be used to determine the effectiveness of the ETD summarisation models. In addition human evaluation will be employed to determine the perceived usefulness of the ETD summarisation models.

The study could potentially provide useful information on the challenges of enabling the general public to engage with ETDs and how the application of NLP and automatic text summarisation can help overcome some of the challenges.

References

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). *Text Summarization Techniques: A Brief Survey* (arXiv:1707.02268). arXiv. http://arxiv.org/abs/1707.02268

Bhandari, M., Gour, P., Ashfaq, A., Liu, P., & Neubig, G. (2020). *Re-evaluating Evaluation in Text Summarization* (arXiv:2010.07100). arXiv. http://arxiv.org/abs/2010.07100

Ermakova, L., Cossu, J. V., & Mothe, J. (2019). A survey on evaluation of summarization methods. *Information Processing & Management*, *56*(5), 1794–1814. https://doi.org/10.1016/j.ipm.2019.04.001

Ingram, W. A., Banerjee, B., & Fox, E. A. (n.d.). *Cadernos BAD Summarizing ETDs with deep learning*.

Johnson, M. E. (n.d.). *Automatic Summarization of Natural Language Literature Review and Synthesis*.

Phiri, L., & M'sendo, R. (n.d.). *Multi-Faceted Automatic Classification of Institutional Repositor Digital Objects*.

Zhang, Y., Wang, R., & Zhou, Z. (n.d.). *Improving Neural Abstractive Summarization via Reinforcement Learning with BERTScore*.