**Universitatea Babeș-Bolyai**

**Facultatea de Matematică și Informatică**

**Departamentul de Informatică**

**A COMPARISON OF SOME DATA-DRIVEN EXTRACTIVE SUMMARISATION STRATEGIES FOR GENERATING ABSTRACTS OF THEOLOGICAL JOURNALS’ ARTICLES**

**Coordonator științific:**

**Lect. univ. dr. Adrian-Ioan Sterca**

**Student:**

**Daniel Avram Pop**

**Cluj-Napoca**

**2021**

The postmodern world we live in today cannot comply with lack of information; as of 2021, around 2.5 quintillions of raw data are produced daily on the internet. Since even trying to mentally grasp this number proves troublesome to most people, a paradoxical situation arises: though access to most of this data is free, one cannot *humanely* benefit from such amounts of bits.[[1]](#footnote-0) Be they the latest news, culinary blogs posts or New Testament exegesis articles, one salient need arises: a manner in which someone can apprehend the most important ideas in a text without having to read it in its entirety. In common parlance, this is known as text *summarisation*.

Though traditionally the task of college students working on their papers and of researchers preparing their work to be published, the development of artificial intelligence, and, more generally, the increase in computers’ computation power made this endeavour suited to be passed on to machines. Thus, with the seminal paper of Luhn[[2]](#footnote-1) in the late fifties, a whole new field was uncovered in the AI: *automatic* text summarisation. But understanding the working structures of human language is no easy venture, since presenting the most relevant parts of a text in a coherent manner supposes both an understanding of the ideas presented and the skill of putting them coherently.[[3]](#footnote-2) Consequently, this and associated tasks fall in the category of the research field called *natural language processing*.

Categorization of automatic summaries has been done in several ways, which we will discuss in due course. Relevant at this moment is the fact that some of the methods are *blind*, ie. they do not take into consideration the type of text they analyze, while others specialize in highlighting the specific features of texts in various domains, like medical papers which have a (relatively) standard structure. What we aim with this study is to analyze several such blind summarisation techniques on very technical pieces of writing: articles from *Novum Testamentum*, a quarterly journal of exegesis published by Brill for more than half a century now, that features such unique features as words in Koine Greek (so non-English script) and in the field of hermeneutics, which are mostly in Latin and German. We hope that our effort will single out whether general automatic summarisation techniques can yield significant results when dealing with such narrow-field texts, and, if that is the case, which are the most relevant *non-*domain specific features of a text.

**Bibliography:**

1. Cer, D., Yang, Y., Kong, S., Hua, N. *et al.*: *Universal Sentence Encoder*, Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Brussels, Belgia, 2018, 169-174.
2. Chatterjee, N., Mittal, A., Goyal, S.: *Single Document Extractive Text Summarization Using Genetic Algorithms*, Third International Conference on Emerging Applications of Information Technology, 2012.
3. Eisele, M.: *Unsupervised Multi-Document Summarization*, 2019, Unpublished master’s thesis, University of Hamburg.
4. Lamsiyah, S., El Mahdaouy, A., Ouatik, S., Espinasse, B.: *A Supervised Method for Extractive Single Document Summarization based on Sentence Embeddings and Neural Networks*, International Conference on Advanced Intelligent Systems, Jul 2019, Marrakech, Moroc.
5. Luhn, H. P.: *The Automatic Creation of Literature Abstracts*, IBM Journal of Research and Development, 2 (1958), p. 159–165.
6. Lin, C.-Y.: ROUGE: *A Package for Automatic Evaluation of Summaries*, Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004), Barcelona, Spania, 2004.
7. Nenkova, A., McKeown, K.: *Automatic Summarization*, The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 5 (2011), 103-233.

1. M. Eisele: *Unsupervised Multi-Document Summarization*, 2019, Unpublished master’s thesis, University of Hamburg, p. 1. [↑](#footnote-ref-0)
2. H. P. Luhn: *The Automatic Creation of Literature Abstracts*, IBM Journal of Research and Development, 2 (1958), p. 159–165. [↑](#footnote-ref-1)
3. Eisele: *Unsupervised Multi-Document Summarization*, p. 1. [↑](#footnote-ref-2)