Advances in Intelligent Systems Research, volume 175
2nd International Conference on Industry 4.0 and Artificial Intelligence (ICIAI 2021)

Malagasy Abstractive Text Summarization Using Scheduled Sampling Model

Volatiana Marielle Ratianantitra[1], Jean Luc Razafindramintsa[1], Thomas Mahatody[1], Claire Rasoamalalavao[1], and Victor Manantsoa[1]

[1] *Laboratory for Mathematical and Computer Applied to the Development Systems, University of Fianarantsoa, Madagascar*
volatianamarielle@yahoo.fr, razafindramintsa.jeanluc@yahoo.fr, tsmahatody@gmail.com, crasoama@gmail.com, vmanantsoa@moov.mg

ABSTRACT
Since 1955, text summarizing has evolved. We could observe all the various approaches in several languages, although most of these methods were for significant languages such as English, French, etc. Other scholars have figured out how to summarize their material in their language (a language other than the major languages), which has led us to discover a way for our language, the Malagasy language, which is considered an under-endowed language. An abstractive text summarizing approach is presented in this study. The abstractive technique is more complicated than the extractive approach because it entails re-formulating the source material while maintaining the general idea. However, it results in a more natural summary and better sentence harmony. The Scheduled Sampling approach was utilized to develop the text summarization model, which used deep learning. The task at hand is to teach the model how to communicate in English. The obtained results suggest that deep learning may be applied to the Malagasy language.

Keywords: *Abstractive Text Summarization, Deep Learning, Malagasy Language, Neuro-Linguistic Programming.*

1. INTRODUCTION

The amount of text being sent and received on the Internet continues to rise at an exponential rate. These data can be condensed while still maintaining its significance and information by summarizing the text. Automatically synthesizing natural language summaries from an input material while retaining the most relevant information is text summarizing. As a result, data might be retrieved more quickly and conveniently. A summation can be either extractive or abstractive, depending on the purpose of the summary.

The extraction method, which is the classic method, was the first to emerge. The primary objective is to find meaningful sentences and include them in summary. It's important to remember that the summary is an exact copy of the original content. The abstractive method, a more advanced strategy, involves recognizing essential areas, understanding the context, and replicating in a new way, utilizing words not found in the original text. Abstract approaches, on the other hand, are more challenging. It's worth noting that summary sentences are generated here rather than just copied from the source material. Deep learning techniques such as the encoder-decoder architecture and LSTM (Long Short-Term Memory) networks are used in this method, which is difficult for beginners to grasp. Because the created sentences aren't present in the original text, the approach can sometimes develop sentences that make no sense. We chose the Malagasy language because there hasn't been much research on it in abstractive text summarization in Natural Language Processing. In addition, Madagascar's national language is Malagasy. Madagascar has some Malagasy dialects. We distinguish between the official Malagasy MO (spoken in the capital) and eighteen regional dialects **[1].** It is important to note that our research focuses on the MO language, which is the most widely used written form in Madagascar. Indeed, the Malagasy language's peculiarities were published in our previous paper **[2].**

The remaining sections of the paper are organized as follows—the related works in Section 2. Section 3 delves into the specifics of the methodology and scheduling approaches. Section 4 discusses the research model and obtained results, followed by Section 5's conclusions. Finally, Section 5 concludes the paper.

2. RELATED WORK

Historically, majority of the work in text summarization has been extractive, which entails selecting sentences in the source document and replicating them as a summary. On the other hand, humans are more likely to generate the original material in their own words. Human summaries are, by definition, abstract. Abstract text summarization has historically been a challenging task for traditional rule-based AI. However, recent breakthroughs in deep learning have permanently altered this. Several abstractive text summarizing studies have used machine learning techniques, which **[3]** summarizes. Inspired by Neural Attention Model's success on the closely related problem of Machine Translation. The mentioned authors in **[4]** that this neural attention model worked extremely well and outperformed earlier non-deep learning algorithms when applied to abstract text summary. However, they limited their analysis to specific sentences and avoided generalization. Indeed, they limited their analysis to single sentences and did not generalize. The authors in **[5]** expanded on this and defined the reference model. In **[6],** the authors have employed pointer generation networks, and the coverage mechanism and **[7]** used a training approach based on reinforcement and intra-attention learning on the decoder outputs. More enhancements were made to this basic model. Both approaches significantly improve rouge scores when compared to the baseline. They re-implemented the latest methodologies and mechanisms used in English on the Arabic language in [**8].** In **[9**], the authors used the Scheduled Sampling model. An Amharic (African language) Dataset that can get good results was mentioned in **[10].** We examine this model in depth before pointing out its flaws. In **[11],** the authors included a Bidirectional GRU with attention for Indonesian text with a source/abstract text ratio of 1.0. The Malagasy language and extracted summary by section 1 are some of the most critical contributions in text synthesis based on deep learning.

As pointed out in **[12],** the concern is that utilizing instructor forcing implies the model has never been trained on its errors and hence may be vulnerable to them—a phenomenon known as exposure bias. When the model is exposed to its own (imperfect) predictions at translation time, this could cause issues.

They were using a planned technique for selecting when to utilize instructor forcing and when not to is a typical approach to tackling the problem of exposure [**13].** Applying planned sampling to a recurrent decoder is simple. For each word generation, the model decides whether to use the gold embedding from the provided target (teacher forcing) or the model prediction from the previous step.

The decoding is still autoregressive in the Transformer model **[14];** however, unlike the RNN decoder, the production of each word is conditional on the entire prefix sequence rather than just the last word. As a result, applying planned sampling to this model is not straightforward. Because the Transformer achieves state-of-the-art results and has become the default choice for many natural language processing tasks, it's worth adopting and exploring the idea of planned sampling for it, which has yet to be offered to our knowledge.

In this study, we make the following contributions:

• We present a new technique for scheduling sampling in seq2seq models that involves performing two passes through the decoder during training.

• We examine numerous ways to conditioning model predictions when employed in place of the gold target.

• We evaluate planned sampling with seq2seq on language pairs in a text summarization task and obtain results comparable to a teacher forcing baseline (with a minor improvement of up to ROUGE and BLEU points).

3. METHODOLOGY

This study's abstractive text summarizes aims to learn how to summaries a succession of phrases in a document. The abstractive text summarization of the Malagasy text was accomplished using a technique. To start building the text summarization model, we'll need a dataset of the required language, Malagasy. The information is gathered from social media (online newspaper articles). It is in CSV format, including texts with titles, to create a template for summarizing the material in the title. It totals around 14k components. We couldn’t accomplish much, but we were able to accomplish something. Then we'll need a language-specific word embedding model to help our machine understand the new language. We will utilize the same genism dataset and Python module for our text-to-text summary model to develop our word embedding model. Then we may create a vocabulary file with all the different terms (150k words), each with its count. To deal with the issue of exposure bias,

we construct a planned sampling model from the attention model (here, we would work on the seq2seq encoder-decoder structure) and the pointer generator model (this is a neural network that is trained to learn when to produce new words, and when to copy words from the original sentence).

The whole summary of the project is given below:
Step 1: Data collection form social media (text with title)
Step 2: Data preprocessing
Step 3: Build word embedding model
Step 4: Vocabulary count in vocab dict
Step 5: Build seq2seq model with scheduled sampling
Step 6: Train model
Step 7: Check the result

4. RESULT AND DISCUSSION

On the Malagasy dataset that we created; we used the scheduled sampling model using python 3. We measure the number of n-grams that overlap between the baseline summary and the summary we made with ROUGE **[14]** and BLEU **[15],** which measure the amount of overlap between the baseline summary and the summary we generated as the measurement grows, suggesting a better result. We performed our assessment on 80 test sentences, the scores were BLEU = 0.2810, ROUGE 1f = 15.49, ROUGE 2f = 04.11, ROUGE Lf = 10.66. *The following* table *gives an example of the result of summary obtained.*

Table 1. Result of training

	Malagasy text
Original text	*Ny soratra, asa sarotra ka tsy voafehy tanteraka raha tsy izarana, ilofosana. Ilàna talenta; kanefa tsy ampy raha tsy tovanana fandalinana maharitra sy fanazaran-tena mitohy. Mila fikarohana ny soratra, mila fanavaozana mandrakariva, ka izay manao azy dia irariana mba hiezaka hatrany, hiezaka tsy miato ka tsy hanadino fa zava-kanto ny soratra ka tokony hitafy endrika manintona.*
Reference	*Asa sarotra ny soratra*
Summary	*Ny soratra tena mila eritreritra*

For comparison, running scheduled sampling model on English, Amharic, and Hindi data sets. On the well-known English CNN / DailyMail dataset, the result was ROUGE-1 of 39.53 and 17.28 ROUGE-2, and on Amharic the scores were BLEU = 0.3311, ROUGE 1f = 20.51, ROUGE 2f = 08.59, ROUGE Lf = 14.76.

For this discrepancy, the English dataset is substantially larger (200k articles with extensive summaries) compared to the African dataset, and it's because English dataset collection is comparatively more accessible. After all, there are so many resources available in this language to draw from.

5. CONCLUSION

The planned sampling technique was applied to the Malagasy language in this study, and we were able to achieve reliable results. This demonstrates that machine learning may be used to process language.

This study will encourage us to reach the final goal, which is to create an abstractive summary that conforms to the type of Malagasy summary, which is to generate extracted sentences (sentences carrying the key ideas) in a form other than the extracted texts. We hope that our work has paved the way for the application of new deep learning techniques to the Malagasy language, as well as the objectives set forth within our contributing research team: research on the processing of the Malagasy language, and the provision of researchers with tools for analysis and automatic processing of the Malagasy language.

REFERENCES

1. Hanitramalala, R.: Vers une typologie des collocations à verbe support en malgache. (2018).
2. Ratianantitra, V. M., Razafindramintsa, J. L., Mahatody, T., & Manantsoa, V.: Deep learning approach for Malagasy text summarization. International Journal of Conceptions on Computing and Information Technology, Vol. 7, Issue. 1, November' 2019; ISSN: 2345 – 9808
3. Sarker, I. H., Kayes, A. S. M., Watters, P.: Effectiveness analysis of machine learning classification models for predicting personalized context aware smartphone usage. J Big Data. 2019; 6:57.
4. Rush, A. M., Chopra, S., Weston, J.: A neural attention model for abstractive sentence summarization. arXiv Prepr arXiv:1509.00685. (2015)
5. Nallapati, R., Zhou, B., Ma, M.: Classify or select: neural architectures for extractive document summarization. arXiv Prepr arXiv:1611.04244. (2016).

6. See, A., Liu, P. J., Manning, C. D.: Get to the point: summarization with pointer generator networks. arXiv Prepr arXiv:1704.04368. (2017)
7. Paulus, R., Xiong, C., Socher, R.: A deep reinforced model for abstractive summarization. In Proceedings of the 2018 International Conference on Learning Representations, (2017)
8. Molham, A. M., Said, D.: Arabic text summarization using deep learning approach. Journal of Big Data, 7(1) (2020).
9. Zaki, A. M., Khalil, M. I., Abbas, H. M.: Amharic Abstractive Text Summarization. arXiv preprint arXiv:2003.13721, (2020).
10. Bengio, S., Vinyals, O., Jaitly, N., & Shazeer, N.: Scheduled sampling for sequence prediction with recurrent neural networks. In Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1, NIPS15, pp. 11711179, Cambridge, MA, USA, MIT Press, (2015).
11. Adelia, R., Suyanto, S., Wisesty, U. N.: Indonesian abstractive text summarization using bidirectional gated recurrent unit. Procedia Computer Science, 157, 581-588, (2019)
12. Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In Advances in Neural Information Processing Systems, pages 1171–1179.
13. Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Niehues Jan, Stuker Sebastian, Sudoh Katsuitho, ¨ Yoshino Koichiro, and Federmann Christian. 2017. Overview of the iwslt 2017 evaluation campaign. In International Workshop on Spoken Language Translation, pages 2–14.
14. Lin, C. Y.: ROUGE: A package for automatic evaluation of summaries. In Text Summarization Branches Out, pp. 74–81, Barcelona, Spain, July 2004.
15. Papineni, K., Roukos, S., Ward, T., Zhu, W. J.: Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002