# IGBOSUM1500 - INTRODUCING THE IGBO TEXT SUMMARIZATION DATASET

**Chinedu Mbonu & Chiamaka Chukwuneke & Roseline Paul & Ikechukwu Onyenwe**
Natural Language Processing and Machine Learning Research Group
Department of Computer Science, Nnamdi Azikiwe University, Nigeria
{ce.mbonu, ci.chukwuneke, ru.paul, ie.onyenwe}@unizik.edu.ng

**Ignatius Ezeani**
UCREL Research Group, School of Computing and Communication
Lancaster University i.ezeani@lancaster.ac.uk

## ABSTRACT

Igbo, like many low-resource African languages, faces the same challenge of inadequate or complete lack of resources - dataset and methods - to support the research and the development of even basic NLP tools for its over 30 millions users. One major gap in the IgboNLP research is the absence of a text summarization tool for Igbo. In this paper, we report our on-going effort in the creation of the `IgboSum1500` dataset, the first standard Igbo text summarization dataset, which will serve as a fundamental precursor to development of the Igbo text summarization resources as well as the expansion of the Igbo and African NLP.

## 1 BACKGROUND

Igbo[1], along with Hausa and Yorùbá, is one of the three prominent indigenous Nigerian languages. It is spoken by the Igbos of southeastern Nigeria with over 30 million speakers resident in Nigeria and many more abroad. In NLP terms, Igbo is still considered to be acutely under-resourced and 'scraping-by' according to Joshi et al. (2020). Currently, efforts are on-going in developing IgboNLP e.g. part-of-speech tagging (Onyenwe et al., 2019), diacritic restoration (Ezeani et al., 2016), embedding based analogy and similarity (Ezeani et al., 2018), machine translation (Ezeani et al., 2020), (Nekoto et al., 2020), and named-entity recognition (Adelani et al., 2021). However, these efforts need to be sustained by creating more resources and expanding the scope of coverage of common downstream NLP tasks in Igbo, and one of such tasks is text summarization.

## 2 OVERVIEW OF TEXT SUMMARIZATION

The growth the internet and web with the accompanying information overload has, among other things, necessitated the quest for tools that can summarise large volumes of texts (Gambhir & Gupta, 2017). The task of text summarization, therefore involves compressing a large body of text into its shorter version which contains only the relevant information in the text (Allahyari et al., 2017).

In the text summarization study, a test dataset is needed to evaluate the performance of any proposed method. Some of the publicly available datasets for the English language summarization tasks include the CNN-Daily Mail dataset (See et al., 2017).

Automatic text summarizers are computer programs that can identify relevant parts of a text document and put them together in coherent and readable way. Common techniques to building automatic text summarizers can be categorized into two key approaches:

- *extractive summarization* (Nallapati et al., 2017; Liu, 2019; Mihalcea, 2005) where the summary contains the exact words and phrases in the original text, and

[1]**Igbo:** https://en.wikipedia.org/wiki/Igbo_language

| **Text:** |
|---|
| *Nkeji edemede 25 nke Nkwupụta Ụwa Nile Maka Ihe Ruuru Ndị Mmadụ nke 1948 nke United Nations na-ekwu, sị: "Onye ọ bụla nwere ikike ibi ndụ zuru oke maka ahụ ike na ọdịmma nke ya na ezinụlọ ya, gụnyere nri, uwe, ụlọ na nlekọta ahụike na ọrụ mmekịrọta dị mkpa". Nkwupụta izugbe gụnyere ebe obibi iji chebe mmadụ ma kwupụtakwa nlekọta enyere ndị nọ n'afọ ime ma ọ bụ nwata. A na-ahụ nkwupụta zuru ụwa ọnụ nke ikike mmadị dị ka nkwupụta mbụ zuru ụwa ọnụ maka oke ikike mmadụ. Kọmishọna Ukwu nke Mba Ndị Dị n'Otu Maka Ihe Ruuru Ndị Mmadụ kwuru na Nkwupụta Ụwa Nile Maka Ihe Ruuru Ndị Mmadụ na-agụnye ọhụụ nke gụnyere ikike mmadụ, obodo, ndọrọ ndọrọ ọchịchị, akụ na ụba,ọha mmadụ ma ọ bụ omenala.* |
| **Reference Summary:** |
| *Nkwupụta Ụwa Nile Maka Ihe Ruuru Ndị Mmadụ na 1948 kwuru na onye ọ bụla nwere ikike ibi ndụ zuru oke. Nke a gụnyere ịnweta nri na uwe na nlekọta ahụike maka onye ọ bụla. Nke a bụ nkwupụta izizi gbasara ikike mmadụ.* |

Table 1: A part of the Igbo version of the UN's Universal Declaration of Human Right showing an example of a reference summary.

- *abstractive summarization* (Gupta & Gupta, 2019; Paulus et al., 2017; Gehrmann et al., 2018) where new words could used, as done by humans, to create the summary

Works on combining the two approached in some hybrid form have also been reported (Chen et al., 2019; Jin et al., 2020; Hsu et al., 2018).

The task of text summarization, which is a form of language understanding and generation, is a complex one. This is because it is hard for the machine to extract the actual meaning of words or phrases in context, inferential interpretation, and generate correct and relevant sentences for the summaries.

## 3 Methodology for Creating 'IgboSum1500'

Given the nature of the source of our data and the time and resources available to the authors, we consider this paper an extended proposal. It details our approach to the core work which is still in progress at the time of submitting this, as well as early evaluations results of the Igbo text summarization systems built.

As shown in Figure 1 we adopted a simple 6-step process for bootstrapping the process of creating the IgboSum1500 dataset as well as baseline Igbo text summarizers which will be discussed in the sections below.

### 3.1 Dataset Collection and Analysis

The main source of our dataset is the website of the Anambra Broadcasting Service[2] - a radio and television station based in one of the major southeastern states Anambra. The choice of this station is mainly because it is the most accessible to the main author who has also secured the permission to use their content.

Although this is a good website for relevant and contemporary local contents across multiple genres, these contents are unfortunately in English. This is challenging given that we aim at building the Igbo text summarization dataset. However, it also provides the opportunity to leverage existing and relatively more developed NLP tools (summarization and translation) for English language in our pipeline for bootstrapping the dataset creation process.

For the purpose of this work we randomly extracted 1500 articles uploaded on the website between the month of May 2021 and February 2022. Figure 2a shows that majority of the articles we collected - over 65% - were published November 2021 and January 2022. We did not investigate whether this

[2] https://www.absradiotv.com/

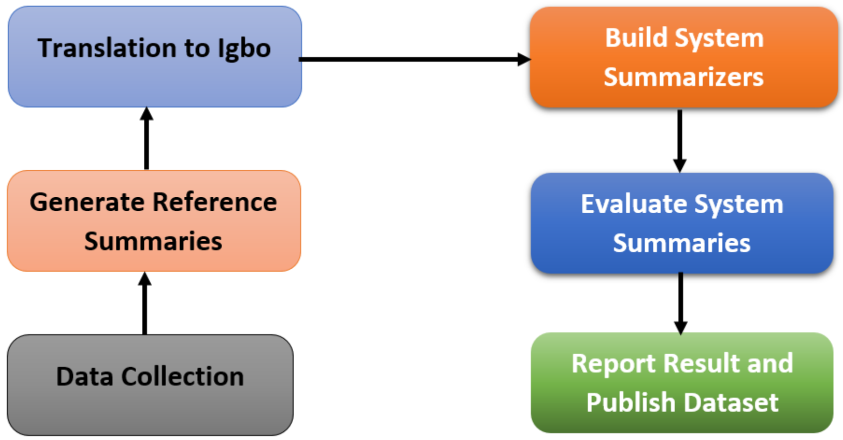


Figure 1: This figure shows the 6-step plan for this dataset creation project with each step discussed in the subsequent sections

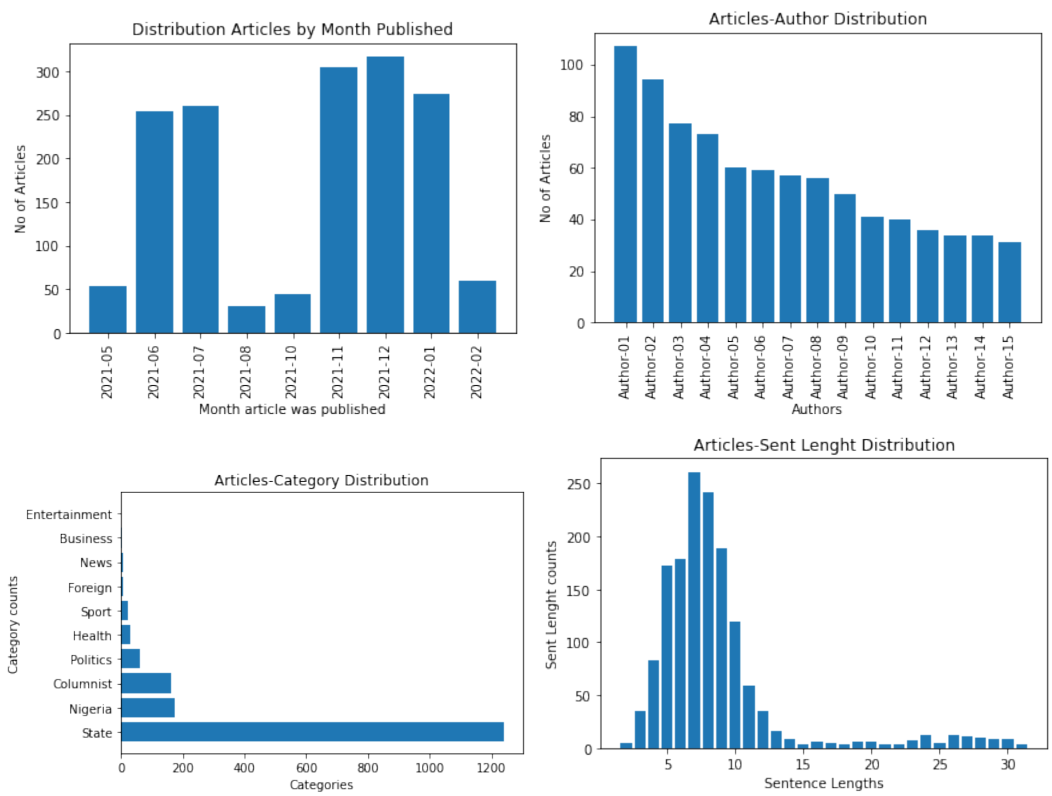


Figure 2: **a:**[top-left] Articles used in this works were published on the website between May 2021 and February 2022; **b:**[top-right] Top-15 authors of the articles extracted; **c.**[bottom-left] Ten articles categories and their distribution; **d.**[bottom-right] Number of sentences in each articles - plot show articles with up to 30 sentences.

is by pure sampling chance or due the electioneering activities happening in the region about the same time.

In total, there are 51 unique authors for all the 1500 articles. However, some of the authors were more prolific than the others leading to only 15 writing over 50% (850) articles. Figure 2b shows an anonymized[3] plot of the counts of articles by each of the top authors.

Looking at categories of the articles as documented on the website, we observed that there were ten unique categories: *Entertainment, Business, News, Sports, Health, Politics, Columnist, Nigeria* and *State*. Each article belongs to at least one of these categories but some articles belong to more than one category. Figure 2c but shows that *State* clearly dominates all the other categories by a large margin.

Another statistic we looked at (Figure 2d) is the number of sentences per article. While there are a lot of articles that are quite long (up to 80 sentences), majority of the articles have between 5 and 10 sentences.

### 3.2 Generating Reference Summaries

The evaluation of text summarization tools is often done by comparing their summaries with gold-standard or 'reference' summaries which are often provided by human summarizers. In order to efficientize this process and given that our original data was in English, we defined a three-stage process (`summarize-evaluate-correct`) that leverages a summarization tool for English as well as Igbo language speakers to create our reference summaries.

The `summarize` stage employs a version of the state-of-the-art BART based (Lewis et al., 2019) summarisation model trained on the `cnn-dailymail` dataset (Nallapati et al., 2016) which is available on HuggingFace[4]. This model was applied to the English version of the articles and then passed on to three language speakers were asked to `evaluate` the quality of the summaries based on some defined criteria:

- **5:** Very clear expression and very readable style. Very few language errors. Relevant knowledge and a good understanding of the article; without significant gaps.
- **4:** Clear expression and legible style. Small number of language errors. Relevant knowledge and a good understanding of the article, with some gaps.
- **3:** Generally clear expression, and legible style. Number of language errors. The knowledge and understanding of the article is sufficient, although there are several omissions and several errors.
- **2:** Expression is generally clear but sometimes unclear. Significant number of language errors. The knowledge and understanding of the article is sufficient for an elementary summary, but there are a number of omissions and errors.
- **1:** Expression is often difficult to understand. Defective style. Persistently serious language errors. The information is inadequate for summary purposes. Obvious deficiencies in understanding the article.

The final stage - the `evaluation` - was actually going on simultaneously with the `correction` stage i.e. the evaluators were instructed to fix errors as they encountered them thereby creating high-quality summaries after the process.

### 3.3 Translating Articles and Summaries

As stated in Section 3.1, our original dataset was in English and so were the generated summaries. However our aim in this work was to build the Igbo text summarization dataset. So having created and corrected the English versions summaries as described in Section 3.2, we proceeded with translating both the articles and their summaries using a standard English-Igbo summarization tool, the GoogleTranslate API[5]. Although this approach was useful for facilitating the process, the quality of the translation was not very good. To improve that, we adopted a similar human-in-the-loop approach to the one in Section 3.2 where language speakers were asked to correct the translated articles and summaries.

[3] We decided to anonymize the authors names to protect their identities.

[4] `https://huggingface.co/sshleifer/distilbart-cnn-12-6`

[5] `https://en.wikipedia.org/wiki/Google_Translate`

### 3.4 Building Summarization Systems

To obtain some initial results, we then built some basic extractive summarisation systems - which will serve as baseline models - and investigated their performances. We built and compared two common extractive summarisation systems - TextRank (Mihalcea & Tarau, 2004) and LexRank (Erkan & Radev, 2004). Both systems use versions of existing ranking algorithms such as PageRank to determine the importance of parts of a text e.g. sentences or phrases.

We used a naive baseline that uses the title of each article as a quasi summary. Some previous works, especially in topic modelling, have noted that similar to the first sentence of a document or the key words, the title of a document does contain some meaningful information about the document (Radev et al., 2004).

### 3.5 Evaluation and Discussion

The system summaries were evaluated by comparing them with reference summarisers using the four commonly used versions of the ROUGE[6] metrics as implemented by Ganesan (2018). *ROUGE-N* (where N= 1 or 2) i.e. unigrams and bigrams; *ROUGE-L* the longest common subsequence in both system and reference summaries that retains the word order; and *ROUGE-SU*: a version of *ROUGE-S*[7] that includes unigrams. ROUGE typically present three key metric scores precision, recall and F1-score as described below.

$$precision = \frac{count(overlapping\ units)}{count(system\ summary\ units)}$$

$$recall = \frac{count(overlapping\ units)}{count(reference\ summary\ units)}$$

$$f1 = (1 + \beta^2) * \frac{recall * precision}{recall + \beta^2 precision}$$

where the value of $\beta$ is used to control the relative importance of *precision* and *recall*. Larger $\beta$ values give more weight to *recall* while $\beta$ values less than 1 give preference to *precision*. In the this work, $\beta$ is set to 1 making it equivalent to the harmonic mean between *precision* and *recall*. The term '*units*' as used in the equation refers to either words or n-grams.

Typically, summarization systems aim at improving the recall score i.e. the fraction of the reference summary it is able to get. Figure 3 shows that the baseline system performed poorly in that regard such that, depsite its high score, the f1 score remained low. This is not surprising given the those summaries were significantly smaller in size that the reference summaries they are comparing against. The other baseline models did much better with significantly higher recall scores mainly because they produced summaries that were of comparable lengths with the reference summaries. Also the higher n-gram scores were generally poor and this is possibly because of their extractive approach to summarisation which is the different from the abstractive and human approach used in creating the reference summaries thereby reducing the chance of longer ngram overlap. TextRank appears to have done slightly better overall given that the f1 scores are higher

## 4 Conclusion and Future work

In this work, we present the first standard, high quality and publicly available Igbo summarisation dataset - IgboSum1500. This is a major contribution to Igbo and AfricanNLP in particular and low-resource NLP in general especially in the natural language understanding and text generation space. We are quite aware of the possible limitations of this work. Using existing tools in other languages may be helpful but may sometimes propagate the errors and biases along the pipeline. Also, the use of basic extractive does not give room for exploration of the challenges text summarisation.

[6]Recall-Oriented Understudy for Gisting Evaluation Lin (2004)

[7]Default *ROUGE-S*: skip-gram co-occurrence of pairs of words in a sentence allowing for arbitrary gaps while maintaining the order

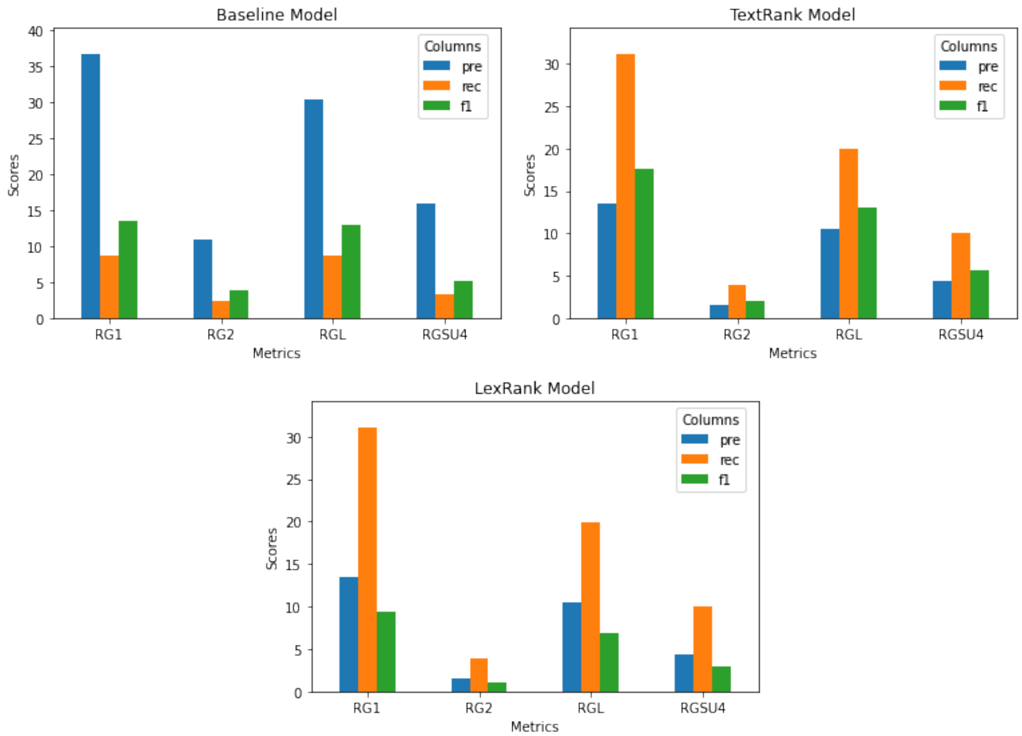


Figure 3: Performance results for the results TextRank and LexRank algorithms compared with the baseline system that uses only the articles titles.

Work is currently ongoing on packaging and releasing the dataset. Future work will focus on experimenting with creating or finetuning state-of-the-art neural models for the Igbo summarisation task focusing on abstractive approach.

## References

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9: 1116–1131, 2021. doi: 10.1162/tacl_a_00416. URL `https://aclanthology.org/2021.tacl-1.66`.

Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, and Krys Kochut. Text summarization techniques: A brief survey, 2017.

Yangbin Chen, Yun Ma, Xudong Mao, and Qing Li. Multi-task learning for abstractive and extractive summarization. *Data Science and Engineering*, 4(1):14–23, 2019.

Günes Erkan and Dragomir R Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479, 2004.

Ignatius Ezeani, Mark Hepple, and Ikechukwu Onyenwe. Automatic restoration of diacritics for igbo language. In *International Conference on Text, Speech, and Dialogue*, pp. 198–205. Springer, 2016.

Ignatius Ezeani, Ikechukwu Onyenwe, and Mark Hepple. Transferred embeddings for igbo similarity, analogy, and diacritic restoration tasks. In *Proceedings of the Third Workshop on Semantic Deep Learning*, pp. 30–38, 2018.

Ignatius Ezeani, Paul Rayson, Ikechukwu Onyenwe, Chinedu Uchechukwu, and Mark Hepple. Igbo-english machine translation: An evaluation benchmark, 2020.

Mahak Gambhir and Vishal Gupta. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, 47(1):1–66, 2017.

Kavita Ganesan. Rouge 2.0: Updated and improved measures for evaluation of summarization tasks. *arXiv preprint arXiv:1803.01937*, 2018.

Sebastian Gehrmann, Yuntian Deng, and Alexander M Rush. Bottom-up abstractive summarization. *arXiv preprint arXiv:1808.10792*, 2018.

Som Gupta and SK Gupta. Abstractive summarization: An overview of the state of the art. *Expert Systems with Applications*, 121:49–65, 2019.

Wan-Ting Hsu, Chieh-Kai Lin, Ming-Ying Lee, Kerui Min, Jing Tang, and Min Sun. A unified model for extractive and abstractive summarization using inconsistency loss. *arXiv preprint arXiv:1805.06266*, 2018.

Hanqi Jin, Tianming Wang, and Xiaojun Wan. Multi-granularity interaction network for extractive and abstractive multi-document summarization. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 6244–6254, 2020.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6282–6293, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.560. URL `https://aclanthology.org/2020.acl-main.560`.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Yang Liu. Fine-tune bert for extractive summarization. *arXiv preprint arXiv:1903.10318*, 2019.

Rada Mihalcea. Language independent extractive summarization. In *ACL*, volume 5, pp. 49–52, 2005.

Rada Mihalcea and Paul Tarau. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pp. 404–411, 2004.

Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*, 2016.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 2144–2160, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.195. URL https://aclanthology.org/2020.findings-emnlp.195.

Ikechukwu E Onyenwe, Mark Hepple, Uchechukwu Chinedu, and Ignatius Ezeani. Toward an effective igbo part-of-speech tagger. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 18(4):1–26, 2019.

Romain Paulus, Caiming Xiong, and Richard Socher. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*, 2017.

Dragomir R Radev, Hongyan Jing, Małgorzata Styś, and Daniel Tam. Centroid-based summarization of multiple documents. *Information Processing & Management*, 40(6):919–938, 2004.

Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. *CoRR*, abs/1704.04368, 2017. URL http://arxiv.org/abs/1704.04368.