Language-based Classification of Words using Deep Learning

Nyambegera Duke Zacharia*, Mwamba Kasongo Dahouda*, Inwhee Joe*
*Dept. of Computer Software Engineering Hanyang University

딥러닝을 이용한 언어별 단어 분류 기법

듀크*, 다후다*, 조인휘*
*한양대학교 컴퓨터 소프트웨어공학과
duke.zacks@gmail.com, dahouda37@hanyang.ac.kr, iwjoe@hanyang.ac.kr

ABSTRACT

One of the elements of technology that has become extremely critical within the field of education today is Deep learning. It has been especially used in the area of natural language processing, with some word-representation vectors playing a critical role. However, some of the low-resource languages, such as Swahili, which is spoken in East and Central Africa, do not fall into this category. Natural Language Processing is a field of artificial intelligence where systems and computational algorithms are built that can automatically understand, analyze, manipulate, and potentially generate human language. After coming to discover that some African languages fail to have a proper representation within language processing, even going so far as to describe them as lower resource languages because of inadequate data for NLP, we decided to study the Swahili language. As it stands currently, language modeling using neural networks requires adequate data to guarantee quality word representation, which is important for natural language processing (NLP) tasks. Most African languages have no data for such processing. The main aim of this project is to recognize and focus on the classification of words in English, Swahili, and Korean with a particular emphasis on the low-resource Swahili language. Finally, we are going to create our own dataset and reprocess the data using Python Script, formulate the syllabic alphabet, and finally develop an English, Swahili, and Korean word analogy dataset.

Keywords: Natural language processing, Python Script, Word representation vectors, Deep Learning, Recurrent Neural network(RNN) Language model, Vocabulary Dataset.

1. INTRODUCTION

We are currently living in an era of advanced technology, meaning nearly every part of our daily lives is related to the technology of some form in one way or another. There is no doubt that over the year's technology has been responsible for creating amazingly useful resources that allow us to have access to the information we need right at our fingertips. The development of technology has led to numerous mind-blowing discoveries, better facilities, and better luxuries, and, at the same time, it has dramatically changed our daily lives in ways unfathomable even just a few years ago. One of the main backbones of the technological development that has taken place recently is that of Artificial intelligence (AI). Inside AI we have machine learning and deep learning. Machine learning (ML) is the scientific study of algorithms and statistical methods that computer systems use to perform a specific task without using explicit instructions and instead rely on patterns and inference.

Deep learning is part of a broader family of machine learning methods based on artificial neural networks. It is a class of machine learning algorithms that uses a cascade of many layers of nonlinear processing. It also part of the broader machine learning field of learning representations of data facilitating end-to-end optimization. It has the ability to learn multiple levels of representations that correspond to hierarchies of concept abstraction [1]. One of the core aspects of such technologic development is Natural language processing (NLP).

Natural language processing (NLP) is a field of artificial intelligence in which computers analyze, understand, and derive meaning from human language in a smart and useful way. Natural language processing (NLP) relies on word embedding as the input for machine learning or deep learning algorithms. The solutions, however, were mainly restricted to machine learning approaches that trained on handcrafted, high dimensional, and sparse features [2]. As time passed, most researchers have now started using neural networks [3], which use dense vector representations. Hence, the superior results of NLP tasks are attributed to word embedding [4,5] and deep learning techniques [6]. As we can see from the authors of [7–10], improved performance of downstream NLP tasks is achieved by learning vector representation of words in language models. Quality word vectors are expected to capture syntactic and semantic similarities among words by addressing the similarities in the surface form of both the words and the context [8]. This has motivated the transition from the conventional one-word

representation to word representation [10] based on words and sub-word information.

However, most of the lower resources languages haven't been given an opportunity to develop in this regard due to the lack of sufficient raw data to work with. In many cases, dealing with low-resource languages requires the ability to deal with raw data supplementation by the researcher. For example, it is quite common to classify words within NLP as somewhere between French and English in many former English colonies in Africa, such as Kenya, Rwanda, Zimbabwe, and South Africa. Thus, expanding the reach of language technologies to users of these languages may require the ability to handle mixed-language data, depending on which domains it is intended for. The purpose of writing and doing the research for this paper is to improve the proposed classification of words in English, Swahili, and Korean languages. According to current research, there have been few, if any, studies done in terms of the classification of words in the Swahili language. This result will be used to improve the Swahili dataset to be recognized in a similar manner to other languages that have already been given priority due to a large amount of data that exists. We shall base this on the use of Python Script and popular network architectures compatible with RNN.

The paper is organized as follows: Section 2 presents related work, and section 3 describes our language modeling architecture and the following section 4 The proposed methods and section 5, Experimental results are presented Finally, in Section 6 will be our conclusion and future works will also be addressed.

2. RELATED WORK

Although NLP has been a major field of research and is well established [11], most of the major problems stem from the fact that the research has focused on high-resource languages. One author [12] has made an attempt to convince the ACL community to prioritize the resolution of the predicaments highlighted here so that no language is left behind. They believe these findings will play a strong role in making the community aware of the gap that needs to be filled before we can truly claim such state-of-the-art technologies to be language agnostic. What they didn't realize, however, is how to properly work with languages without resources of data and they didn't know if a universal method could be applied to all languages.

Author [13] worked on the Swahili syllabic alphabet and the word analogy dataset for Swahili, but the limitation he faced was a lack of proper data, and he was forced to use the English dataset provided by Mikolov et al. [14] to build the Swahili dataset. Although work has been put into the study of Natural Language Processing (NLP) for Swahili word classification, the results are insignificant. Most of the researchers, such as the above authors, have done research on Swahili using the existing English dataset, whereby there is no accurate result due to inappropriate data. That's why we decided to create our own dataset from scratch so as to run the model dataset and find an accurate value and result using the Recurrent Neural Network Language Model (RNN).

3. LANGUAGE MODELING

Language modeling (LM) is the use of various statistical and probabilistic techniques to determine the probability of a given sequence of words occurring in a sentence. Language models analyze bodies of text data to provide a basis for their word predictions. They are used in Natural Language Processing (NLP) applications, particularly ones that generate text as an output. Some of these applications include machine translation and question answering. They interpret this data by feeding it through an algorithm that establishes rules for context in natural language. It is a crucial element in modern NLP applications because it is the reason that machines can understand qualitative information. Each language model type, in one way or another, turns qualitative information into quantitative information. It is used directly in a variety of industries, including tech, finance, healthcare, transportation, law, the military, and government. Additionally, it's likely that most people reading this have interacted with a language model in one way or another at some point in the day, whether it be through a Google search, and autocomplete text function, or interaction with a voice assistant. The start of all of this was when the author Markov [12] wrote his paper.

Machine learning and deep learning algorithms have been instrumental in NLP [13] with word embedding playing an important role in the success of the language. These algorithms widely depend on the availability of data. If we look at the current situation today, we can find that a lack of NLP resources and systems has been a major limitation for NLP development and achievements, which has led to the necessity of terms such as high-resource languages and low-resource languages. As such, we see that many low-resource languages, such as Swahili and African tribal, have inadequate resources and systems available. This is the main challenge when confronting the issue of how to do a project without a dataset. Therefore, in this article, we shall contribute resources that will support English, Swahili, and Korean language modeling, as well as lay a foundation framework for other NLP tasks.

4. PROPOSED METHOD

1) Dataset description

Due to the lack of data provided for lower resource languages, such as Swahili, we decided to create our own dataset from scratch to help us get an accurate result. Our data is generated from day-to-day conversation words, plus words from Swahili, English, and Korean dictionaries.

	noun	Language
0	forms	ENGLISH
1	able	ENGLISH
2	about	ENGLISH
1709	uangaze	SWAHILI
1710	udongo	SWAHILI
1711	ufumbuzi	SWAHILI
2709	지불	KOREAN
2710	지상	KOREAN
2711	지수	KOREAN

Firstly, we had to input the data and it was then loaded as an Excel dataset to our python script. We executed it as shown in the figure above. Our dataset is composed of three kinds of different languages. Secondly, we had to encode the dataset because most of the machine learning algorithms cannot operate on label data directly. They require all input variables and output variables to be numeric. This is commonly required due to the constraint of the efficient implementation of machine learning algorithms rather than hard limitations on the algorithms themselves. Thus we had to convert the categorical dataset into a numerical form as shown below.

	noun	Language_Encoded
0	forms	0
1	able	0
2	about	0
562	hivyo	1
563	hofu	1
564	hoja	1
2495	물	2
2496	오리	2
2497	오프	2

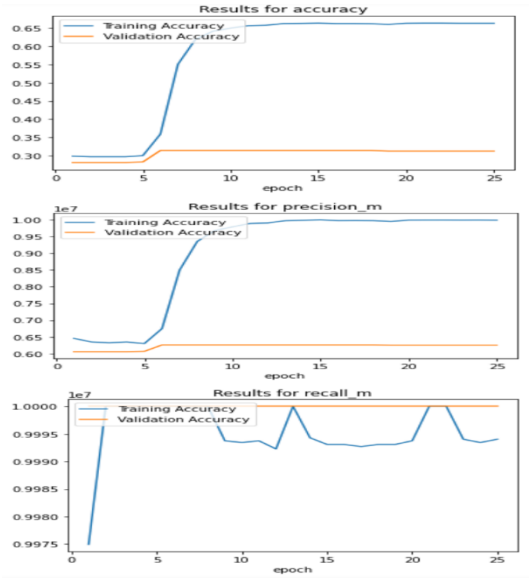
Then after completing all of the above, we had to explore the data to check how many 0,1 and 2's we have in the dataset and we also needed to check the missing data to be sure that all our data was up to date. The next step, which is important for the data that we corrected, is tokenization. Tokenization is the process of splitting up a larger body of text into smaller lines, words or even creating words for a non-English language. While creating and training a Machine Learning or Deep Learning model we need to make the model understand the given text. So, in order to accomplish this purpose, we needed to tokenize the texts. During the process of interpreting the data from the model, all inputs and outputs were tokenized in the same way as done while training the model. That's why we hard to tokenize our data so that we could create the model easily.

2) **Recurrent neural network**.

A recurrent neural network (RNN) is a type of artificial neural network which uses sequential data or time-series data. These deep learning algorithms are commonly used for ordinal or temporal problems, such as language translation, Natural Language Processing (NLP), speech recognition, and image captioning. In a similar vein as feedforward and convolutional neural networks (CNNs), recurrent neural networks utilize training data to learn. They are distinguished by their "memory" as they take information from prior inputs to influence the current input and output. While traditional deep neural networks assume that inputs and outputs are independent of each other, the output of recurrent neural networks depends on the prior elements within the sequence. We trained the RNN model to be able to read the dataset and to create the dataset in a way where we could recognize and classify the words of the three languages that we wanted to learn through the model. It can be considered a major issue that Swahili remains as one of the low resources languages without a dataset and has not been given more opportunity by researchers despite it being spoken by more than 16 million people as a dialect and more than 32 million as a second language. Although it was simple enough to find an English dataset and a Korean dataset, we decided to create our dataset from scratch whereby the data would be composed of words that could be categorized as nouns, verbs, and adjectives. This data will help us to classify the words we need into Korean, English, or Swahili and later use the same data for speech recognition.

5. EXPERIMENT & RESULTS

The aim of this paper is the classification of words into English, Swahili, and Korean while primarily focusing on the lower resource language of Swahili. For our experiment, we had to run the dataset that we made on the python script. We created the dataset and then cleaned up the dataset utilizing the following steps. At the start, due to working with different languages, we were required to label our data with unique values. Therefore, we gave English a value of 0, Swahili a value of 1, and Korean a value of 2 respectively, and we removed the punctuation and stop words as well as finally going through the process to tokenize the dataset. We loaded our dataset after cleaning it, trained and test set it. Also, we trained the tokenizer and used that tokenizer to convert the sentence into sequences of numbers for it to be able to work with our mode. Then, we padded the sequences so that each sequence could be viewed the same at each length. After going through all of the above steps, we built our model framework. Then, we compiled the model so that it could fit the Recurrent Neural Network. We plotted the evaluation metrics across epoch so as recurrent neural network model (RNN) to evaluate the performance on the test dataset. The accuracy, precision, recall, and loss function predicted values for training and validation accuracy and testing stages computed by various methods are displayed below.



6. CONCLUSION AND FUTURE WORK

We have presented the raw Swahili, English, and Korean dataset and even used the dataset to classify the word representation vectors. Our studies focused on machine learning in terms of natural processing language and python to compose and learn the data. We demonstrated the quality of the word embedding by building a language model that could help us to have a comprehensive, competitive, and perfect result of the dataset. We also confirmed the quality of RNN using the word analogy task after developing the Swahili, English, and Korean analogy dataset. The performance of the model depends on the quality and the quantity of its respective word representation. We, therefore, propose using RNN for the classification and recognition of words in different languages, machine translation, and text classification tasks in future works. Furthermore, I will consider and focus on working in the future for the development of this paper by delving further into advances CNN for the improvement of the dataset that will thus improve the word embedding and classification of the African triple languages.

REFERENCES

1. Y. Bengio, A. Courville, P. Vincent, "Representation Learning: A Review and New Perspectives". IEEE Transactions on Pattern Analysis and Machine Intelligence. 35 (8): 1798–1828., 2013, doi:10.1561/2200000006.{ https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5816885/}
2. Young, T.; Hazarika, D.; Poria, S.; Cambria, E. Recent trends in deep learning based natural language
3. Processing. IEEE Comput. Intell. Mag. **2018**, 13, 55–75. [CrossRef]
4. McCulloch, W.S.; Pitts, W. A logical calculus of the ideas immanent in nervous activity. Bull. Math. Biophys.**1943**, 5, 115–133. [CrossRef]
5. Mikolov, T.; Karafiát, M.; Burget, L.; C˘ ernocky` , J.; Khudanpur, S. Recurrent neural network based language model. In Proceedings of the Eleventh Annual Conference of the International Speech Communication Association, Chiba, Japan, 26–30 September 2010.
6. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems; Neural Information Processing Systems Foundation, Inc.: Cambridge, MA, USA, 2013; pp. 3111–3119.
7. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. Nature **2015**, 521, 436. [CrossRef] [PubMed]
8. Collobert, R.; Weston, J. A unified architecture for natural language processing: Deep neural networks with multitask learning. In Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland, 5–9 July 2008; pp. 160–167.
9. Botha, J.; Blunsom, P. Compositional morphology for word representations and language modelling. In Proceedings of the International Conference on Machine Learning, Beijing, China, 21–26 January 2014; pp. 1899–1907.
10. 8. Ling, W.; Dyer, C.; Black, A.W.; Trancoso, I.; Fermandez, R.; Amir, S.; Marujo, L.; Luís, T. Finding Function in Form: Compositional Character Models for Open VocabularyWord Representation. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; Association for Computational Linguistics: Stroudsburg, PA, USA, 2015; pp. 1520–1530. [CrossRef]
11. A Dataset for Examining the State of NLP Research https://www.aclweb.org/anthology/2020.lrec-1.109/
12. Pratik Joshi_ Sebastin Santy_ Amar Budhiraja_ Kalika Bali Monojit Choudhury {The State and Fate of Linguistic Diversity and Inclusion in the NLPWorld}
13. Casper S. Shikali ,Zhou Sijie, Liu Qihe , and Refuoe Mokhosi,::Better Word Representation Vectors Using Syllabic Alphabet: A Case Study of Swahili
14. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation ofWord Representations in Vector Space. In Proceedings of the 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, AZ,USA, 2–4 May 2013.
15. A Mathematical Theory of Communication. https://ieeexplore.ieee.org/document/6773024
16. Word-Level Language Identification and Predicting Codeswitching Points in Swahili-English Language Data (https://www.aclweb.org/anthology/W16-5803.pdf)