IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Issue 5, May 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105004 |

Part of Speech Tagging for Shona

Paul Confidence Chikovo[1], Greenford Walter Katuruza[2], Raymond Zenda[3], L Nhapi[4], Pianos Gweme[5]

Software Engineering Student, Dept. of SE., SIST, Harare Institute of Technology, Harare Zimbabwe[1]

Software Engineering Student, Dept. of SE., SIST, Harare Institute of Technology, Harare Zimbabwe[2]

Software Engineering Student, Dept. of SE., SIST, Harare Institute of Technology, Harare Zimbabwe[3]

Software Engineering Lecturer, Dept. of SE., SIST, Harare Institute of Technology, Harare Zimbabwe[4]

Software Engineering Student, Dept. of SE., SIST, Harare Institute of Technology, Harare Zimbabwe[5]

ABSTRACT: The written form of Shona, the most widely used language in Zimbabwe, presents unique difficulties for NLP. Speech tagging, or the practice of giving textual words tags, is a fundamental NLP activity. These word labels (tags) can be straightforward like a noun, verb, or adjective or as intricate as a singular noun or a third-person past tense verb. Although POS tagging is not helpful on its own, it is widely acknowledged as the initial step in comprehending a natural language. The difficulty of POS labelling is due to ambiguity. All languages contain ambiguity in some way. Clarifying ambiguities demands precise and effective approaches. In order to enhance POS tagging performance for Shona, theoretical and practical POS tagging challenges have been studied in this dissertation. It has been determined that understanding Shona structure, the provided corpora, and the labelling methods all have significant factors in the performance outcome. employing the most recent machine learning algorithms for the studies. This much enhanced performance can be attributed to the usage of a partly tagged dataset, appropriate feature selection emerging from structural understanding ofthe language,andparameter adjustments derivingfrom expertise of thetagging methods

KEYWORDS: Brill, Shona, NLTK, HMM, POS, TnT, SVM, CRF, tag

I. INTRODUCTION

Human linguistic information and knowledge are growing abundant in a world that is becoming a one village. The more technology develops, the more each natural (human) language and culture interact with one another. More than ever, there is a need to develop and advance natural language technologies.The process of developing software tools that enable computers to understand human languages is known as artificial intelligence (AI). NLP is a component of AI. NLP can be applied at several levels, such as the word, phrase, sentence, or semantic levels[1].

It is evident that computers cannot comprehend human languages the same way that people do. They are unable to perceive the meanings and syntax of words in a phrase. However, given that each natural language's data grows, it becomes more challenging for humans to manually analyse and extract the necessary components from it [2]. To control the vast amount of data already there, we require the assistance of computers. Natural language processing has become a fascinating field of computer science as a result of this need for computers' assistance.

Part-of-speech (POS) tagging, is a fundamental task in natural language processing. It involves giving each word in a text a tag, such as a noun, pronoun, verb, preposition, adverb, adjective, or other vocabulary class identifier[1]. Although POS tagging is not helpful on its own, it is usually acknowledged as the initial step in comprehending a natural language. It is significantly reliant on for the majority of other jobs and applications.

Words' parts of speech are frequently confused. For instance, the English word "store" might be an infinitive, a finite verb, or a noun. This ambiguity in a sentence is typically resolved by the word's context. For instance, the word store can only be an intuition in the sentence "The 1977 P6's could store two pages of data." A system known as a part-of-speech tagger uses contextual information to automatically classify words' parts of speech. There are numerous potential uses for part-of-speech taggers, including speech synthesis, machine translation, information retrieval, and speech recognition[4].

II RELATED WORK

English, Tonga, Ndebele, and many more languages are members of the family of structured languages, which also includes Shona. These languages have a number of things in common, therefore the knowledge and outcomes of NLP

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105004 |

activities performed on one of them can be applied to the others. When it comes to POS labeling,these languages have similar issues.

Their writing methods and nature of word construction lead to the first shared issue. For a very long time, the principal structured languages have existed as written languages. Their writing systems support the use of white space to separate words.The words in these languages, however, are not the same as those in English. They are created by joining together lexical components, the majority of which may belong to different word classes (POS). In structured languages, two or more words in English can be regarded as one word.

Because of this, the POS tagging procedure for these languages is more difficult. The POS tagging units are not immediately obvious because words are made up of several different morphemes with potential boundary ambiguities. Should the morphemes or the phrases as they exist in text (separated by a space) be POS tagged? This question's answer concludes both the tag-set design and the tokenisation technique.

English has the greatest reported tagging accuracy rates thus far. In truth, POS tagging is typically regarded as a problem that has been resolved for English. Accuracy levels have gotten to about 97%. There is an unmet demand for new or modified approaches, particularly for languages with limited resources. This explains in part why so much recent research on POS tagging has been devoted to creating new methods for these languages or modifying those that already exist.

IIISHONA MORPHOLOGY

The Shona alphabet has 23 basic characters. Each such character is modified in some regular fashion to reflect the five vowels of the language. Therefore, there are in total $23 * 5 = 115$ characters[38]

$$C = \{A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,R,S,T,U,V,W,Y,Z\}$$

The vowels are:

$$V = \{a,e,i,o,u\}$$

According to Fortune (1955) [40], there are two basic conjugations of the Shona verb: affirmative and negative. The moods imperative, indicative, prospective, participial, relative, subjunctive, and hortative are further divided into for each conjugation. Additionally, four the future, the present, the immediate past, and the distant past. According to aspect, mode, and implication, tenses are separated. The characteristics are unbounded, ongoing, and flawless.

To make the analysis more understandable, I'll use the verb -enda, which means "go," as an example to concentrate on the indicative mood of the affirmative conjugation. The fundamental prefixes of the Shona verb are revealed by examining the indefinite simple, which is described in Table 1.

Tenses	Incidental	Habitual
PRES	Ndi-no-enda ‘I will go’	Ndi-no-enda ‘I usually go’
PAST	Nda-enda ‘I went’	Nda-i-enda ‘I used to go’
RM.P	Nda-ka-enda ‘I went’	-----
TURE	Ndi-cha-enda	-----

The accidental present form is created when the prefix ndi- joins with the suffix -no. The prefix -cha-, which comes after the prefix ndi-, denotes the incidental's future tense. The past tenses are distinguished by a second set of subject markers with the vowel a serving as a defining trait. In order to create the incidental recent past, the prefix nda- is thus immediately applied to the root. After the past person marker, the prefix -ka- designates the distant past. There are only two tenses in the habitual mode: present and past (the future and remote past occur in the incidental only). Another

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105004 |

challenge that can be observed from Shona words is that the same word can have more than two meanings, a good example for this is the word zuva which can mean day or the sun.

IV THE TAGGING PROCESS

Several different open source implementations of machine learning techniques, including CRF++,LIBSVM, and NLTK, are used in the studies in this dissertation. Conditional Random Fields (CRFs) are a C++ open source library that are used to segment and label sequential data. LIBSVM implements support vector-based classification and regression algorithms in C++ and Java. It has connections with numerous languages. We conducted our trials using the Python interface that LIBSVM provides (via SWIG). The Natural Language Toolkit, sometimes known as NLTK, is a collection of open source software modules that covers both conceptual and empirical naturallanguageprocessing techniques. It issimple to use, learn, and customize. In this dissertation, POS tagging experiments have been conducted using NLTK versions of TnT and Brill[36]. Jupyter notebook which is found in Anaconda will be used for the tagging process.

The tagging procedure involves the following tasks:

- Tokenisation
- Feature Extraction
- Disambiguation

Step 1: Tokenisation

In this step sentences are broken down to identify the basic units of a language so that they can be given tags. An example from the lyrics of the song from Winky D "[Imi vakuru woye gadzirai ramangwana ramangwana revana, haridi makwati kuritambanudza ramangwana revana. Haridi hurombwa kuritambanudza ramangwana revana. Tese tavakunge mangererere tsika dzemangerere. Tirikuchema nhamo tirikuchema hurwere, tipei nhoroondo izere. Tarisai wetsvimbo makore aenda asi achingori mugota. Vakuwasha vodzosa vanasikana hanzi gupuro tichazopa]"can be split using the *nltk.word_tokenize()*function to produce the following as shown in the image below:

```
In [22]: from nltk.tokenize import word_tokenize, sent_tokenize
         z = nltk.word_tokenize(x)
         y = nltk.sent_tokenize(x)
         print(z)
         ##print(y)

['Imi', 'vakuru', 'woye', 'gadzirai', 'ramangwana', 'ramangwana', 'revan
a', ',', 'haridi', 'makwati', 'kuritambanudza', 'ramangwana', 'revana',
'.', 'Haridi', 'hurombwa', 'kuritambanudza', 'ramangwana', 'revana', '.',
'Tese', 'tavakunge', 'mangererere', 'tsika', 'dzemangerere', '.', 'Tiriku
chema', 'nhamo', 'tirikuchema', 'hurwere', ',', 'tipei', 'nhoroondo', 'iz
ere', '.', 'Tarisai', 'wetsvimbo', 'makore', 'aenda', 'asi', 'achingori',
'mugota', '.', 'Vakuwasha', 'vodzosa', 'vanasikana', 'hanzi', 'gupuro',
'tichazopa']
```

Figure 1: Tokenisation

Step 2: Feature Extraction

After tokenization, a given unit must have its word and context information extracted in a way that classification algorithms may use them. In contrast to English, Shona words contain a lot of information regarding syntax [16].

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105004 |

However, the placement of words in a phrase can provide a lot of useful information. For instance, unless the text is a poetry, the primary verb in Shona is always at the end of the phrase. Nouns come before adjectives. Verbs are placed after adverbs.

Step 3: Disambiguation

The next step is to choose the appropriate tag for a word based on its characteristics. Assigning the appropriate class label to a given input is the task of classification in machine learning [17]

```
In [17]: nltk.pos_tag(z)

Out[17]: [('Ngano', 'NNP'),
          ('dzevaShona', 'VBZ'),
          ('inzira', 'JJ'),
          ('yekufambiswa', 'NN'),
          ('kwemashoko', 'NN'),
          (',', ','),
          ('nemafungiro', 'RB'),
          (',', ','),
          ('nemaitiro', 'RB'),
          (',', ','),
          ('nekuvaraidza', 'JJ'),
          ('uyewo', 'JJ'),
          ('nekudzidzisa', 'JJ'),
          ('kwakadzama', 'NN'),
          ('nekurongeka', 'NN'),
          ('.', '.'),
          ('Zvinoreva', 'NNP'),
          ('kuti', 'VBD'),
          ('ngano', 'JJ'),
          ('isimba', 'NN'),
          ('guru', 'NN'),
          ('rekuvaka', 'NN'),
          ('munhu', 'NN'),
          (',', ','),
          ('misha', 'NN'),
          ('nenyika', 'FW'),
          ('.', '.'),
          ('Kutangisa', 'NNP'),
```

Figure 2: Tagged text

II. CONCLUSION AND FUTURE WORK

The Since there wasn't a morphological analyser available until recently, Shona POS tagging efforts have thus far relied on supervised stochastic algorithms with annotated data [18]. Itwillbe fascinating to see how themorphologicalanalysercanbeincluded into thesetagging approaches now that one is available Exploring arule-basedtagger basedon thesame or a related morphologicalanalyser will also be interesting. On the basis of the findings outlined in this section, the following potential new works can be suggested [19].

- This study on the Shona language employed a medium-sized, lower-quality corpus as their dataset. The dataset was taken from a PHD student who wrote the thesis in Shona. It can be quite vital to work on enhancing the quality and size of the current corpus [19].
- The Shona POS tagger can be improved by separating the associated prepositions and conjunctions with the other parts of speech while preserving the important information giving morphology like gender, person, number, etc. The morphologies denoting gender, person, number, and other information are ignored when dividing prepositions and conjunctions. Such significant morphologies are ignored when determining the split preposition or conjunction and the root word. It can be crucial to change the segmentation process so that certain morphologies are preserved.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105004 |

- A hybrid Shona POS tagger with features created manually and by neural word embedding techniques may also perform better.

REFERENCES

[1] J. P. Ferraro, H. D. Iii, S. L. Duvall, W. W. Chapman, H. Harkema, and P. J. Haug, "Improving performance of natural language processing part-of-speech tagging on clinical narratives through domain adaptation," pp. 931–939, 2013, doi: 10.1136/amiajnl-2012-001453.

[2] N. K. S. K. N. Malarvizhi, "Bi - directional LSTM – CNN Combined method for Sentiment Analysis in Part of Speech Tagging (PoS)," *Int. J. Speech Technol.*, no. 2017, 2020, doi: 10.1007/s10772-020-09716-9.

[3] D. Das, S. Petrov, R. Mcdonald, and J. Nivre, "Token and Type Constraints for Cross-Lingual Part-of-Speech Tagging," vol. 1, pp. 1–12, 2013.

[4] C. Lv, H. Liu, Y. Dong, and Y. Chen, "Corpus based part-of-speech tagging," *Int. J. Speech Technol.*, vol. 19, no. 3, pp. 647–654, 2016, doi: 10.1007/s10772-016-9356-2.

[5] W. B. Demilie, "Parts of Speech Tagger for Awngi Language," vol. 9, no. 9, 2019, [Online]. Available: http://ijesc.org/

[6] E. Chabata, "The Shona Corpus and the Problem," vol. 10, no. July 1999, pp. 75–85, 2000.

[7] D. Abuzeina, W. A. Moustafa, and H. Al-muhtaseb, "Toward enhanced Arabic speech recognition using part of speech tagging," pp. 419–426, 2011, doi: 10.1007/s10772-011-9121-5.

[8] M. Argaw, "Amharic Parts-of-Speech Tagger using Neural Word Embeddings as Features," *Master's thesis*, 2019.

[9] K. Stratos, M. Collins, and D. Hsu, "Unsupervised Part-Of-Speech Tagging with Anchor Hidden Markov Models," vol. 4, pp. 245–257, 2016.

[10] T. Dalai, T. K. Mishra, and P. K. Sa, "Part-of-Speech Tagging of Odia Language Using statistical and Deep Learning-Based Approaches," vol. 1, no. 1, 2022, [Online]. Available: http://arxiv.org/abs/2207.03256

[11] M. Yasunaga, J. Kasai, and D. Radev, "Robust Multilingual Part-of-Speech Tagging via Adversarial Training," 2016.

[12] D. Grantor, "Tihana Britvi ´ c SEMI-SUPERVISED NEURAL PART-OF-SPEECH TAGGING Master thesis," 2022.

[13] B. Gupta Banik and S. K. Bandyopadhyay, "Novel Text Steganography Using Natural Language Processing and Part-of-Speech Tagging," *IETE J. Res.*, vol. 66, no. 3, pp. 384–395, 2020, doi: 10.1080/03772063.2018.1491807.

[14] C. Lv, H. Liu, Y. Dong, F. Li, and Y. Liang, "Using Uniform-Design GEP for Part-of-Speech Tagging ¤," vol. 26, no. 4, pp. 1–14, 2017, doi: 10.1142/S0218126617500608.

[15] S. Alam, A. Sushmit, Z. Abdullah, and S. Nakkhatra, "Bengali Common Voice Speech Dataset for Automatic Speech," no. 11.

[16] X. Xue, S. M. Asce, J. Zhang, D. Ph, and A. M. Asce, "Building Codes Part-of-Speech Tagging Performance Improvement by Error-Driven Transformational Rules," vol. 34, no. Liao 2005, pp. 1–10, 2020, doi: 10.1061/(ASCE)CP.1943-5487.0000917.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105004 |

[17] N. Entities, “Sequence Labeling for Part of Speech and Named Entities Part of Speech Tagging”.

[18] G. Lkhagvasuren, J. Rentsendorj, and O. E. Namsrai, “Mongolian Part-of-Speech Tagging with Neural Networks,” *Smart Innov. Syst. Technol.*, vol. 212, pp. 109–115, 2021, doi: 10.1007/978-981-33-6757-9_15.

[19] A. P. Herrero, P. F. Lamas, and J. M. R. Presedo, “Parameterization of state duration in Hidden semi-Markov Models: an application in electrocardiography,” pp. 1–9, 2022, [Online]. Available: http://arxiv.org/abs/2211.09478

BIOGRAPHY

Paul Confidence Chikovois a Masters student in the Software Engineering Department, Harare Institute of Technology. Paul received his Bachelor of Science Honours in Computer Science in 2017 from National University of Science and Technology, Bulawayo, Zimbabwe. His research interests are in Natural Language Processing, recommender systems, and Deep learning.

Greenford Walter Katuruza is a Software Engineer at Zimbabwe Centre for High performance Computing (ZCHPC) and Space Scientist at Zimbabwe National Geospatial and Space Agency (ZINGSA). He is currently pursuing Master of technology at Harare Institute of Technology, School of Information Science and Technology, Department of Software Engineering, He received Bachelors of Science Honours Geography in Geospatial Intelligence degree in 2019 from University of Zimbabwe, Harare, Zimbabwe. His research interests are on Deep Learning, AI, Algorithms, and Automation, Space technology etc.

Raymond Zenda is a Masters student in the Software Engineering Department, School of Information Science and Technology, Harare Institute of Technology (HIT). He received Bachelor of Technology degree in Information Technology degree in 2019 from HIT, Harare, Zimbabwe. His research interests are Deep Learning, Big Data Analytics, Natural Language Processing etc

Mr L Nhapi is a lecturer in the Software Engineering department, School of information Science and Technology (HIT). He is interested in Neural Networks, Recommender systems and Data structures

Pianos Gweme is a spatial data scientists at the Zimbabwe National Geospatial and Space Agency with vast experience in near real time spatial data collection using UAVs,mobile devices, satellites etc for various applications such as agriculture, mining, disease surveillance, environmental management etc. He is the chief investigator of ZIMSAT-1 the first satellite of Zimbabwe and Chairman of Technical committee of the Zimbabwe center for high performance computing. His research interests are in remote sensing, satellite communication, UAVs software architecture and precision agriculture using UAVs.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.379

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details