KAMBA PART OF SPEECH TAGGER USING MEMORY BASED APPROACH.

Benson N. kituku[1], Musumba George[1]andPeter Wagacha[2]

[1] Department of Computer Science, DedanKimathi University of Technology, Nyeri, Kenya.
[2]School of computing and informatics, University of Nairobi ,Nairobi, Kenya.

ABSTRACT

Part of speech tagging is very important and the initial work towards machine translation and text manipulation. Though much has been done in this regard to the Indo- European and Asiatic languages, development of part of speech tagging tools for African languages is wanting. As a result, these languages are classified as under resourced languages.

This paper presents data driven part of speech tagging tools for kikamba which is an under resourced language spoken mostly in Machakos, Makueni and Kitui. The tool is made using the lazy learner called Memory Based Tagger (MBT) with approximately thirty thousand word corpuses. The corpus is collected, cleaned and formatted with regard to MBT and experiment run.

Very encouraging performance is reported despite little amount of corpus, which clearly shows that using the state of art technology of data driven methods tools can be developed for under resourced languages. We report a precision of 83%, recall of 72% and F-score of 75% and in terms of accuracy for the known and unknown words, and accuracy of 94.65% and71.93% respectively with overall accuracy of 90.68%..This predicts that with little source of corpus using data driven approach, we can generate tools for the under resourced languages in Kenya.

KEYWORDS

Part of speech tagging, Natural language processing, k-fold, F-score and Ambitag

1.INTRODUCTION

Evolution of the internet and computing gadgets has made a lot of data and information available (offline and online) in different languages. To analyze the data and information automatically for consumptionpurposes [1], then Natural Language Processing Tools (NLP) are needed.Of over the 7000 living languages [2] in the world, only Indo-European and Asian languages have perfect working tools. However, for under resourced languages [3], [4] and [5] mostlyfound in Africa, few have NLP tools. Hence, in this paper, we explain how to make Part of Speech tagger for Kikamba which is an under resourced language in Kenya

Part of Speech tagging (PoS) is the process in which syntactic categories are assigned to words or mapping from sentences to strings of tags [6].Part of speech tagging is also known as word classes, morphological classes, or lexical tags. Example of part of speech in Kikamba include: Noun like Nyumba, verb like enda, adjective like Nzeo etc.PoS provide a lot of information about the word itself and its neighborshence task (PoS) is of key usefulness to many subsequent

DOI : 10.5121/ijnlc.2015.4204

manipulations of text, in that it provides a useful abstraction from the actual words if we want to process all words that belong to special class (get all verbs in a documents) and also provides superficial degree of disambiguationat the different levels of processing. For example, parsing or on itself, let's consider pronunciation of the word 'discount'. If it exists as a noun, we would emphasize on the *dis* during pronunciation, that is:'DIScount' while if it is in the form of a verb we would emphasize on *count*, that is:disCOUNT hence ease the work of text to speech conversion [7]. Other areas where PoS is highly used is in information retrieval, speech synthesis and recognition, question answering and machine translation

Kikamba (Kamba) is a Bantu language spoken by almost four million Kamba people in Kenya, according to the 2009 population & housing census [8] and is the 5th largest indigenous language in Kenya according to the table 2.1 below. Most of this population lives in

No	LANGUAGE	SPEAKERS
1	KIKUYU	6,622,576
2	LUHYA	5,338,666
3	KALENJIN	4,967,328
4	LUO	4,044,440
5	KAMBA	3,893,157
6	KENYAN SOMALI	2,385,572

Table 1 Samples of language speakers in Kenya

Machakos, Makueni andKitui counties, and a substantial number along the Embu, TaitaTaveta and Tharaka boundaries. They are known to have migrated to Kenya from Congo through Tanzania before settling in their current places. Their main economic activities are crop and livestock farming, bee keeping, carving and basketry. For a long time the Kamba people have been known for their culture through carving, especially at Wamunyu, also basketry (kiondo) and traditional dance (kilumi). In 2005 the Akamba Culture Trust (ACT) [1]was formed with the agenda of crusading for the preservation of culture through the written form literature among others. Despite the efforts, unique skills and the number of people speaking the language in the organization, there is still very minimal corpus available online (digital text) and there is very little commercial interest in the languages thus the language still remains in the class of under resourced languages.

2.BACKGROUND

In Kenya, Part-of-Speech tagging has been investigated in two languages namely Swahili and Gikuyu. Hurskainen[9] has extensively researched the Swahili PoS using Finite state methods while [10] has used machine learning methods. De Pauw et al. [11]has done some PoS preliminary experiments for Dholuo language and finally, using machine learning methods,De Pauw[12] developed one for the Gikuyu language. Outside Kenya, some work has been done for the South African languagefor example a number of tag sets and preliminary systems are available for Xhosa [13], and Northern Sotho [14][15].

[1] http://www.machakos.org/Concept_Paper.pdf

2.1 Approaches to building PoS Tagger

2.1.1 Rule based/grammar based approach

Mainly in terms of structure consist of two parts namely: a dictionary and rules [7].The Dictionary contains words and their possible part of speech tagswhile the rules help to disambiguate where more than one part of speech is assigned to a word. The rules may include things to do with morphology etc. and usually they are of two forms: lexical rules and contextual rules [16]. Though the approach provides high accuracy, it requires a lot of man power to create the rule and one need to be aware of the linguistics feature of the language in question.

2.1.2 Maximum entropy model (MEM)

Make use of several observations from the inputs but one is taken at a time, then extraction of useful features from the single observation of a word is done, Finally based on the extracted features you classify the word to the tag set with the highest probability [7].

2.1.3 Brill tagging

Introduced by brill in 1994 [17] and also called transformation based learning, It uses rules which are generated from data (corpus) by machine learning techniques automatically. The rules are learned by the following stages in reference to [7]

- Words in corpus are assigned most suitable tag sets
- Select the maximum tagging based on all possible transformation
- Using established rule, the data is re-tagged again
- Then you repeat the last two until no more improvement

2.1.4 HMM (Hidden Markov model)

Probabilities is the key engine of this methods, usually uses Bayesian inference model [7]. Therefore, the part of speech tagging is treated as a classification problem. The tokenized words are given has a sequence of observations to the classified, then prediction of the class of the tag set. The prediction of the class uses the product of prior probability of the tag set and maximum likelihood of the tag set [7]

2.1.5Memory based Learning.

Daelemans [6] introduces the concept similar to cased based reasoning, where training is done to a classifier and the results of the tag set stored in the memory. Then by use of machine learning algorithm similarity is done between the new inputs and what is stored in the memory and based on the similarity metrics, classification to the right tag set can be done. This is a data based approach and the one used to build the Kamba tagger.

2.1.6 Genetics tagging.

Ali[16] in 2013 introduced this evolutionary model of approach by doing some experiment on Arabic language which searches optimal solution of the tag set by use of heuristics means. Makes use of three operators: fitness operator, mutation operator and cross operator to make the classification.

2.1.7Hybrid.

The model tends to mix rule based approach and data based approach. Rule based approaches achieves high precision while data based approaches achieves high coverage. The aim is to try to strike a balance between precision and coverage

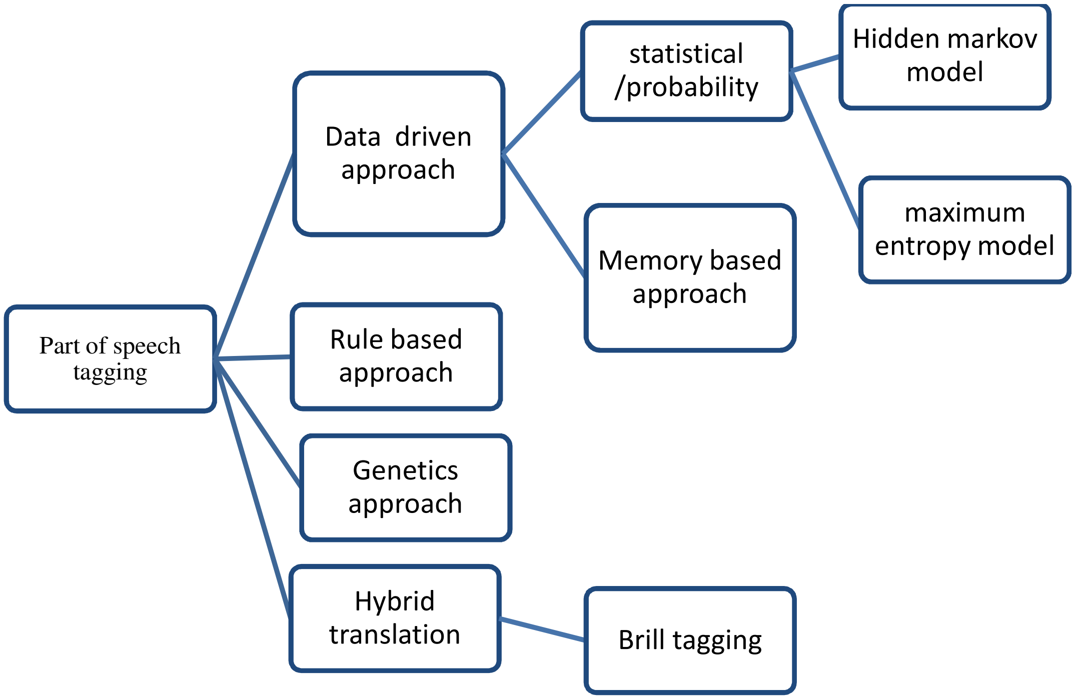


Diagram 1 summary of part of speech tagging approaches

3.ROADMAP TO BUILDING THE POS TAGGER

The data approach methods was used to develop the tagger since rule based involves hand crafting a dictionary of parts of speech and an extensive list of hand-written disambiguation rules which is time consuming, particularly for an under resourced language such as Kikambathere are no known expert of the language. We have employed the second route for building the Kikamba part of speech tagger. We used a lazy learning tagging tool called memory based tagger (Mbt) [18]. The diagrams 2 represent the architecture of building the whole PoS system.

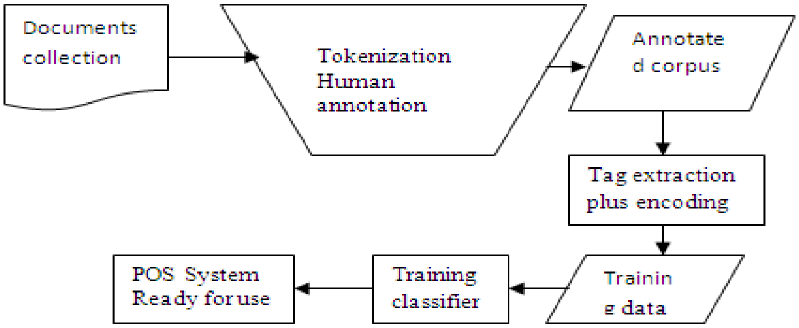


Diagram 2. Architecture of the POS system

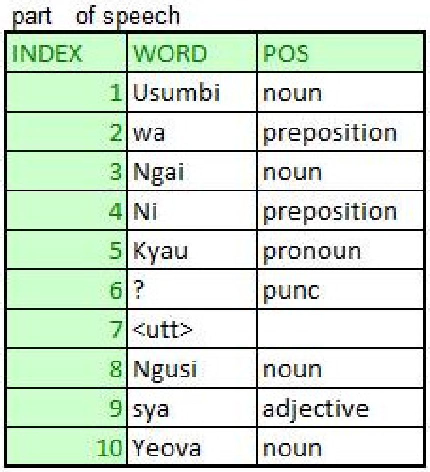
3.1 Corpus collection

Corpus was collected from online and some documents written in Kikamba language. To increase the reliability and validity of the tagger, output cleaning of the corpus was done. Some of the impurities to be purged included but are not limited to wrong punctuation, wrongly inserted words and corrupted information. After this tokenization which involves separating words and punctuation, the corpus was changed to a two column format as required by the Memory Based Tagger (MBT[2]) tool. One column is the word and the other column is the word tag. Microsoft excel was used to help in formatting one word per cell in one column, sentence after sentence up to the end. Approximate thirty thousand words were used in this experiment of which 2000 thousand were generated manually to help diversify the scope of the corpus which was basically a religious corpus.

3.2 Annotation

The manual annotation of the word tags was done in the Microsoft excels. For the purpose of this experiment the PoS tags used were: adjective, noun, preposition, verb, pronoun, Interjection, number, adverb, punctuation and conjunction. However, some of them were abbreviated e.g. Num for Number. The annotated template was transferred into text file still maintaining the format, and run on the MBT tag that was in Linux operating system. To show the end of a sentence,the MBT tagger uses the symbol <utt>. Below is an extract of annotation

[2] Mbt_3.1_manual.pdf page 10



part of speech

INDEX	WORD	POS
1	Usumbi	noun
2	wa	preposition
3	Ngai	noun
4	Ni	preposition
5	Kyau	pronoun
6	?	punc
7	<utt>	
8	Ngusi	noun
9	sya	adjective
10	Yeova	noun

Table 2.0 Annotation

3.3 Tagger generation:

The ready annotated corpus is run on MBT. Firstly, a frequency lexicon tags file is created, where the contents are words with different possible tags as have been assigned in the corpus plus the frequency of occurrence of each in the corpus, resulting in what is referred to as Ambitag[3](An ambitag is a symbolic label defining for a word the different tags it can have according to the corpus). For example, the word ‘discount’ in English language can have the tag of verb and noun and then its frequency in the corpus counted.

The MBT tagger has two smaller taggers; known and unknown word taggers, which are generated after the frequency tags. The former classifies user input which is already in the corpus while the latter classifies new words which donot exist in the corpus and need more information to be disambiguated. The two are the ones used to classify any input text. For each to be generated during training period, different foci were used which are stated below.

For known words group:

- Focus on two disambiguated words on left.
- Focus on one ambitguation tag on the right plus the word being disimbiguated.
- Focus for each context tags two words on the left for the corresponding word.
- Focus for each context tag one word on the right for the corresponding word.

For unknown words group:

- Focus one disimbiguated tag on the left
- Focus one ambiguation tag to the right
- Focus on the first two letters of the word to be taggede.g with Mb, Nd, Ng, Ny, Th.
- Focus on the last two letters of the words to tagged
- Focus on the capitilazation of the word

[3] Mbt_3_1_Manual.pdf page 6

- Focus on the hyphens. E.g Ng'eng'eta, Ng'ombe, Ng'ota

For unknown words, a bit of morphological structure was used so as to help us gather more information about the word to be tagged and predict correctly the right tag.

4. RESULTS AND EVALUATION

4.1 Experimental set up

Weiss [19] k-folds model of testing system was employed because of the small amount of corpus that was used. K-fold model involves partitioning the corpus into K equal portions. The partitioning should be done at the sentence boundary so as to ensure the contexts of words are not affected. Then 90% of the K-folds are used as the training data and the rest as the testing data. For our case, the K was 10, each part consisting of approximately 3000 words. Hence ten runs were made, each time changing the test portion. Overall metrics were gotten by averaging the ten runs.

4.2 Evaluation metrics:

For the purpose of evaluation, four metrics were used to judge the performance of the classifier namely: recall, Precision, F-score and accuracy. According to Walter [20], while performing classification processes let us say with class C, the following subclasses occur: The true positives cell (TP) contains a count of examples that have class C and are predicted to have this class correctly by the classifier. The false positives cell (FP) contains a count of examples of a different class that the classifier incorrectly classifies as C. The false negatives cell (FN) contains examples of class C for which the classifier predicted a different class label than C and true negatives cell (TN) contains a count of examples that are not of class C but the classifier predicted them to be of this class C.

Precision is the proportional number of times that the classifier has correctly made decision.Some instances are in class C. The proportional number of times the classifier assigns class C of test data instances is called recall while the weighted harmonic mean of recall and precision is called F-score and all of them are given by the formulae 1,2 and 3 respectively:

$$Precision = \frac{TP}{TP+FP} \text{ Formula 1}$$

$$Recall = \frac{TP}{TP+FN} \text{ Formula 2}$$

$$F\_score = \frac{2.precision*recall}{precision+recall} \text{Formula 3}$$

4.3 Results

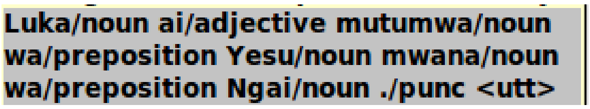
The experiment run reported a precision of 83%, a recall of 72% and an F-score of 75%. The table 3.0 below depicts the precision, recall and F-score of each class or category of what tag was being tested. The accuracy of known and unknown tags was 94.65% and71.93% respectively with overall accuracy of 90.68%.

class |	**precision**	**Recall**	**F-score**
noun |	0.77679	0.97072	0.86269
preposition |	0.94701	0.94282	0.94456
pronoun |	0.98277	0.85629	0.91394
punc |	0.99705	0.98282	0.98978
adjective |	0.95677	0.90407	0.92912
conjuction |	0.84605	0.93234	0.88623
verb |	0.86144	0.41123	0.55465
interjection |	0.8556	0.78332	0.81175
adverb |	0.94955	0.84857	0.89449
num |	0.98309	0.32485	0.47663
exclamation	0	0	0
AVERAGE	**83.24%**	**72.34%**	**75.13%**

Table 3.0 F-score

4.4 Test run

The classifier been ready for other use, people were able to test other sentence. Below is a result of the sentence that was inputted to the classifier "***Luka aimutumwawaYesuMwanawaNgai "***. The sentence in English is translated " Luke was an apostle of Jesus the son of God" The classified gave an output as shown in figure below



```
Luka/noun ai/adjective mutumwa/noun
wa/preposition Yesu/noun mwana/noun
wa/preposition Ngai/noun ./punc <utt>
```

Only the word "ai" has been misclassified as adjective while it's a verb, the rest are correctly classified in their right morph- categories.This is clear indication the classifier was working well, many sentences were also tested.

4.5 Discussion

The precision of 83% with an approximately thirty thousand word corpus is very encouraging. However,a closer look at each of the categories as per the table 3.0 reveals that nouns contributed to the low performance with a performance of 77%. The rest of these categories were above the average performance. The implication according to formula 1 is that many other categories were classified as nouns resulting in an increase of the false positive (FP) sub class. A close analysis of the noun misclassified on each run according to confusion matrix extracted for the MBT tagger can be represented by figure 1 and shows every run with at least 1000 nouns.Approximately 200 nouns were misclassified.

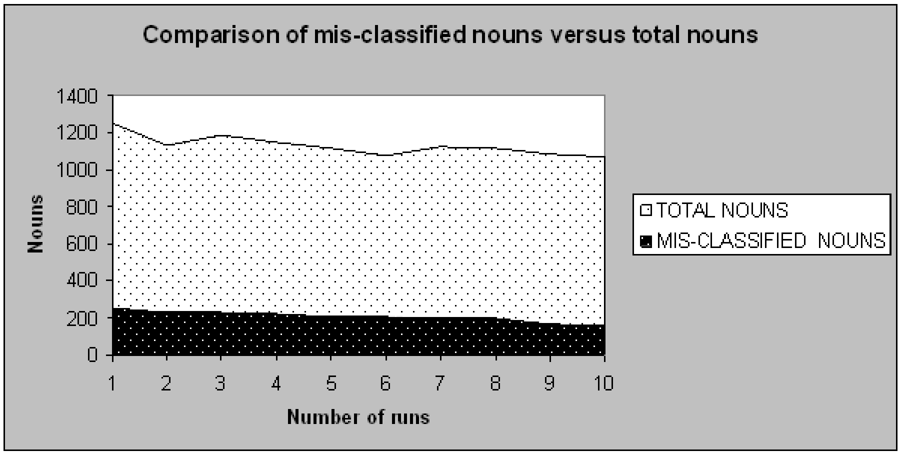


Figure1.Mis-classified nouns versus total nouns

De Pauw[11]carried out the triangular part of speech tagging experiment of English- Dholuo and Swahili - Dholuo with over 100 thousand word corpus resulting in projection precision of 69.7% and 68.4%, hence despite low corpus, the kikamba part of speech tagging performed very well. The verb and numbers which had a recall of 41% and 32% respectively brought down the recall to 72% ,meaning other categories were classified as verbs and numbers other than their correct classes resulting in too many false negatives(FN). The experiment shows an average of 75% as F-score which is good performance.

In terms of accuracies though too below the onefor the Swahili part of speech tagging that had accuracy for known and unknown words of 98.43% and 89.81% respectively and overall accuracy of 98.81 %[10]. However, the Swahili which also used memory based tagging development tool may have performed better because it had five times the corpus used for this experiment. The triangular experiment on English- Dholuo and Swahili - Dholuo reported a projection accuracy of 51.3% and 48.9% though this is understandable because the first step was translation from either English or Swahili to Dholuo and then assigning a tag which might have contributed to the overall low performance.

On coverage, an experiment was done by starting with 2000 words and continued to increase the words by 2000 and noting the accuracy which is as tabulated below. Every increase in corpus resulted to increase of accuracy but a break even resulted at 18000 words, thus it's clear more corpus would result to high accuracy

numbers of words	accuracy overall
2k	77.77
4k	79.26
6k	81.15
8k	81.98
10k	83.64
12k	84.53
14k	85.25
16k	85.58
18k	86.14
20k	96.01

Table 4. Accuracy versus total numbers of words

5. CONCLUSION

The part of speech tagging presented in this paper with precision, F-score and accuracy of 83%, 75% and 90.6% respectively is clear evidence that data driven methods can be used to make tools for under resourced languages such as Kikamba with little corpus. However, to enhance the accuracy,more corpus can be passed through the tagger and resulting tags checked by human experts and rectified where possible, and the resulting annotated corpus be put as part of the training corpus.

For purpose of future work I do propose development of Kikamba machine translator which use part of speech as one of the engine.

REFERENCE:

[1] Lopez, A. (2008). Statistical machine translation.ACM Computing Surveys (CSUR), 40(3), pg 8.

[2] Gordon, R.G.J., (ed.),(2005). Ethnologue: Languages of the World, Fifteenth edition. Dallas,Tex.: SIL International.

[3] Berment V., (2004). “Méthodes pour informatiser des langues et des groupes de languespeudotées” PhD Thesis, J. Fourier University – Grenoble I, May 2004.

[4] Muhirwe, J. (2007). Towards Human Language Technologies for Under-resourced languages.In Joseph Kizza et al. (ed.) Series in Computing and ICT Research, 2Fountain Publishers, 123-128.

[5] De Pauw, G. &Wagacha, P.W. (2007).Bootstrapping morphological analysis of Gikuyu using unsupervised maximum entropy learning. In H. Van hamme& R. van Son (Eds.), Proceedings of the Eighth Annual Conference of the International Speech Communication Association. Antwerp, Belgium.

[6] Daelemans, W. Zavrel, J.van der Sloot, V and van den Bosch, (2002).“MBT: Memory- Based Tagger, version 1.0, Reference Guide,” ILK Technical Report ILK-0209, University of Tilburg, The Netherlands. 2002.

[7] Jurafsky,D& James H. Martin, (2006). Speech and Language Processing: An introduction to natural language processing,computational linguistics, and speech recognition. Copyright 2006.

[8] Oparanya,W,A . (2010). 2009 Population & Housing Census Results.Available from http://www.knbs.or.ke/Census Results/Presentation by Minister for Planning revised.pdf], Nairobi, Kenya.

[9] Hurskainen, A (2004). HCS 2004 – Helsinki Corpus of Swahili Compilers: Institute for Asian and African Studies (University of Helsinki) and CSC

[10] De Pauw,G and De Schryver, G.M and Wagacha,P,W (2006). ("Data-Driven Part-of-Speech Tagging of Kiswahili," in Proceedings of Text, Speech and Dialogue, 9th International Conference, vol 4188, Lecture Notes in Computer Science,P. Sojka, I. Kopecek, K. Pala, Eds. Verlag:Springer, 2006.

[11] De Pauw, G., Maajabu, N.J.A. &Wagacha, P.W. (2010).A knowledge-light approach to Luo machine translation and part-of-speech tagging.In G. De Pauw, H. Groenewald& G-M de Schryver (Eds.), Proceedings of the Second Workshop on African Language Technology (AfLaT 2010). Valletta, Malta: European Language Resources Association (ELRA), pp. 15–20.

[12] De Pauw, G., De Schryver, G. M., & van de Loo, J. (2012).Resource-light Bantu part-of-speech tagging.Language Technology for Normalisation of Less-Resourced Languages, 85.

[13] Allwood,J. Grönqvist, L and Hendrikse, A ,P (2003). Developing a tagset and tagger for the African languages of South Africa with special reference to Xhosa. Southern African Linguistics and Applied Language Studies, 21(4):223–237.

[14] Prinsloo,J and Heid, U(2005). Creating word class tagged corpora for Northern Sotho by linguistically informed bootstrapping. In Proceedings of the Conference on Lesser Used Languages & Computer Linguistics (LULCL-2005), Bozen/Bolzano, Italy

[15] Faaß,G. (2010). The verbal phrase of northern sotho: A morpho-syntactic perspective. In G. De Pauw, H.J. Groenewald, and G-M.deSchryver, editors, Proceedings of the Second Workshop on African Language Technology (AfLaT 2010), pages 37–42, Valletta, Malta. European Language Resources Association (ELRA).

[16] Ali,B and Jarray,F (2013). Genetics approach for arabic part of speech tagging, international journal of Natural language computing (IJNLC) volume 2.

[17] E.Brill (1995). Transformation based error driven learning and natural language processing.A case study in part of speech tagging. Compuational linguistics 21(4) pg 543-566

[18] Daelemans, W, Zavrel, J. Van den Bosch, A. and Van der Sloot, K. 2007. Reference Guide (15 pages, 100 kB PDF); BT: Memory-Based Tagger, version 3.1, Reference Guide. ILK Technical Report Series 07-08.

[19] Weiss, S., &Kulikowski, C. (1991). Computer systems that learn. San Mateo, CA:Morgan Kaufmann

[20] Zavrel, J. and Daelemans, W. 1997. Memory-based learning: Using similarity for smoothing.In Cohen, P. R. and Wahlster, W., editors, Proceedings of the Thirty-Fifth Annual Meeting of them Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics, pages 436.443, Somerset, New Jersey. Association for Computational Linguistics.

Author

Benson Kituku

He is a lecturer at Dedan Kimathi University of Technology in the department of computer science, his under graduate and Masters are in computer science at Maseno and Nairobi universities respectively. He has several years of industrial experience. Currently pursuing PhD in Computer science at university of Nairobi (Machine translation). My major fields of interest for research work are Natural language processing, Automata theory And compiler construction.

George Musumba

He is a Lecturer and PhD student atDedan Kimathi University of technology, Nyeri kenya

Peter Wagacha

He is an associate professor at school of computing and informatic at the university of Nairobi,Kenya