Proceedings of the First workshop on Resources for African Indigenous Languages (RAIL), pages 21–24
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Complex Setswana Parts of Speech Tagging

Malema, G. Tebalo, B. Okgetheng, B. Motlhanka, M. Rammidi G.
University of Botswana
P/Bag 704, Gaborone, Botswana
{malemag,rammidig}@ub.ac.bw, {bokgetheng, mofenyimoffat}@gmail.com

Abstract

Setswana language is one of the Bantu languages written disjunctively. Some of its parts of speech such as qualificatives and some adverbs are made up of multiple words. That is, the part of speech is made up of a group of words. The disjunctive style of writing poses a challenge when a sentence is tokenized or when tagging. A few studies have been done on identification of multi-word parts of speech. In this study we go further to tokenize complex parts of speech which are formed by extending basic forms of multi-word parts of speech. The parts of speech are extended by recursively concatenating more parts of speech to a basic form of parts of speech. We developed rules for building complex relative parts of speech. A morphological analyzer and Python NLTK are used to tag individual words and basic forms of multi-word parts of speech respectively. Developed rules are then used to identify complex parts of speech. Results from a 300 sentence text files give a performance of 74%. The tagger fails when it encounters expansion rules not implemented and when tagging by the morphological analyzer is incorrect.

Keywords: parts of speech tagging, Setswana, qualificatives

1. Introduction

Setswana is a Bantu language spoken by about 4.4 million people in Southern Africa covering Botswana, where it is the national and majority language, Namibia, Zimbabwe and South Africa. The majority of speakers, about 3.6 million, live in South Africa, where the language is officially recognized. Setswana is closely related to Southern and Northern Sotho languages spoken in South Africa. There have been few attempts in the development of Setswana language processing tools such as part of speech tagger, spell checkers, grammar checkers and machine translation.

Setswana like other languages is faced with ambiguity problems as far as word usage is concerned and this has much impact in part of speech (POS) tagging. Text in the available resource Setswana corpus is not annotated and hence limited meaningful processing can be executed on the data in its current form. Setswana as a low resourced language is limiting corpus research pertaining to much needed significant amount of information about a word and its neighbours, useful in further development of other applications such as information retrieval, collocation and frequency analysis, machine translation and speech synthesis, among other NLP applications. Therefore, there is need to develop basic Setswana processing tools that are accurate and usable to other systems.

Parts of speech tagging identifies parts of speech for a given language in a sentence. The output of a POS tagger is used for other application such as machine learning, grammar checking and also for language analysis. The complexity of POS tagging varies from language to language. There are different approaches to part of speech tagging, the most prominent being statistical and rule-based approaches (Brants 2000, Brill 1992,1995 and Charniak 1997). Statistical approaches require test data to learn words formations and order in a language. They work well where adequate training data is readily available. We could not find a readily available tagged corpus to use. We therefore developed a rule based approach. Rule based techniques require the development of rules based on the language structure.

Setswana like some Bantu languages is written disjunctively. That is, words that together play a particular function in a sentence are written separately. For the sentence to be properly analysed such words have to be grouped together to give the intended meaning. There are several orthographic words in Setswana such as concords which alone do not have meaning but with other words they give the sentence its intended meaning. Some of these words also play multiple roles in sentences and are frequently used. Such 'words' includes include *a, le, ba , se, lo, mo, ga, fa , ka*. Without grouping the words, some words could be classified in multiple categories. This problem has been looked at as a tokenization problem in some studies (Faaß et al 2009, Pretorius et al 2009 and Talajard and Bosch 2006)

Setswana parts of speech include verbs, nouns, qualificatives, adverbs and pronouns. Verbs and nouns are open classes and could take several forms. Studies in parts of speech tagging have concentrated on copulative and auxiliary verbs and nouns because of this (Faaß et al 2009 and Pretorius et al 2009). Most of POS taggers have focused on tagging individual words. However, Setswana has POS in particular qualificatives and some adverbs that are made up of several words and in some cases about a dozen words (Cole 1955, Mogapi 1998 and Malema et al 2017). Setswana qualificatives include possessives, adjectives, relatives, enumeratives and quantitatives. Adverbs are of time, manner and location.

A few studies have been done on tokenization and parts of speech tagging for Setswana and Northern Sotho which is closely related to Setswana (Faaß et al 2009 and Malema et al 2017). These studies have not covered tokenization of complex parts of speech. Adverbs, possessives and relatives have a recursive structure which allows them to be extended resulting in complex structures containing several POS. Complex in this case, we mean in terms of length and use of multiple POS to build one part of speech.

This paper investigates identification of Setswana complex qualificatives and adverbs using part of speech tagger. We present basic rules on how to identify complex POS such as adverbs, possessives and relatives. The proposed method tags single words and then builds complex tags based on developed expansion rules. The rules have been tested for

relatives and preliminary results show that most rules are consistent and work most of the time.

2. Setswana Complex POS

As stated above adverbs, possessives and relatives have a recursive structure that allow a simple POS to be extended into a complex POS. We have noted that in Setswana sentence structures, the verb can be followed by noun (object) or an adverb as also stated in the structure of Setswana noun and verb phrases (Letsholo and Matlhaku 2014). We have also noted that nouns could be followed by qualificatives. Thus a simple sentence could be expanded by stating the object the verb is acting on and how, where and when the verb action is performed. The object could be described by using qualificatives and demonstratives.
Examples:

mosimane o a kgweetsa (the boy is driving)
mosimane o kgweetsa koloi (the boy is driving a car)
mosimane o kgweetsa koloi ya rraagwe (the boy is driving his father's car)
mosimane o kgweetsa koloi ya rraagwe kwa tirong (the boy is driving his father's car at work)

The first sentence does not have an object. In the second sentence an object (*koloi/car*) is provided for the verb *kgweetsa(drive)*. In the third sentence, the object *koloi* is distinguished or modified by using the possessive '*ya rraagwe' (his father's).* In the fourth sentence an adverb of place *(kwa tirong/at work)* is added to identify where the action (*kgweetsa/drive*) of driving is happening.
We have observed that possessives, relatives and adverbs have a recursive structure and therefore could be expanded using other POS to create a complex POS.

2.1 Possessives

Simple possessives are made up a concord followed by a noun, pronoun or demonstrative.

Examples:

kgomo ya kgosi (chief's cow)
kgomo ya bone (their cow)

In the first example above "*ya kgosi*" is the possessive, where *ya* is the possessive concord matching the noun class (class 9) of *kgomo(cow)* and *kgosi(chief)* is the root (noun in this case).
This is the simplest form of the possessive. However, the root can be expanded to form a complex possessive. The root can be other compound POS such as relatives, possessives, adjectives and adverbs. These compound roots could also be expanded using the sentence expansion rules as explained above. That is, if POS ends with a verb, the verb can be given an object and or an adverb in front of it and if the POS ends with a noun, the noun can be modified with a qualificative and or a demonstrative. The added POS could also be expanded in the same way recursively.
Examples:

koloi ya monna (the man's car)
koloi ya ntate yo o thudileng (the car that belongs to the man who had an accident)
koloi ya ntate yo o thudileng tonki (the car that belongs to the man who hit a donkey)
koloi ya ntate yo o thudileng tonki ya kgosi (the car that belongs to the man who hit the chief's donkey)
koloi ya ntate yo o thudileng tonki ya kgosi kwa morakeng (the car that belongs to the man who hit the chief's donkey at the cattle post)

In the first sentence *ya monna*, is just the possessive concord and a simple root (*monna/noun*). The second sentence expands the possessive by distinguishing the *monna(man)* with the relative, *yo o thudileng*. Since that relative ends with a verb we could give it an object, *tonki (donkey)* as done in the third sentence. The fourth sentence distinguishes the *donkey(tonki)* using another possessive, *ya kgosi*. The firth sentence adds an adverb of place, *kwa morakeng (at the cattle post)*, for the verb *thudileng (hit).* Further expansion of the possessive could be done by providing objects and or adverbs for new verbs and modifying new nouns with qualificatives or demonstratives. In the last sentence "*ya ntate yo o thudileng tonki ya kgosi kwa morakeng*" is a possessive describing the noun *koloi*. This possessive is made up of a relative (*yo o thudileng*), noun (*tonki*), possessive (*ya kgosi*) and adverb (*kwa morakeng*). The main objective of this study is to develop ways to recognize such long/complex parts of speech.

2.2 Relatives

Relatives are made up of a concord and a root.
Example:

koloi e e thudileng (the car that had an accident)

e e thudileng is a relative, where *e e* is a relative concord for class 4 and 9 nouns and *thudileng* is the root. Using the same approach for expansion of verbs and noun we could expand this relative as follows.

koloi e e thudileng tonki (the car that hit a donkey)
koloi e e thudielng tonki ya kgosi (the car that hit the chief's donkey)
koloi e e thudileng tonki ya kgosi kwa morakeng (the car that hit the chief's car at the cattle post)

In the third example "*e e thudileng tonki ya kgosi kwa morakeng*" is a relative made up of a basic relative (e e *thudileng*), noun (*tonki*), possessive (*ya kgosi*) and adverb (*kwa morakeng*). The structure of examples above is referred to as direct relatives. Another category of relatives is known as indirect. Examples:

koloi e ba e ratang (the car they like)
koloi e a tla e rekang (the car she/he will buy)
ngwana yo Modimo a mo segofaditseng (the child that God blessed)
koloi

In this study we only looked at direct relatives which have a simpler structure compared to indirect relatives. Basic structures of Setswana qualificatives and adverbs could be found in (Cole 1955 and Mogapi 1998).

2.3 Adverbs

Adverbs could also be expanded when they use verbs and nouns. Examples:

kwa morakeng (at the cattle post)
kwa morakeng wa monna (at the man's cattle post)
kwa morakeng wa monna yo o berekang (at the cattle

post of the man who is working)
kwa morakeng wa monna yo o berekang kwa sepateleng (at the cattle post of the man who is working at the hospital)
kwa morakeng wa monna yo o berekang kwa sepateleng sa Gaborone (at the cattle post of the man who is working at Gaborone hospital)

The last example is an adverb made up of a basic adverb (*kwa morakeng*), possessive (*wa monna*), relative (*yo o berekang*), adverb (*kwa sepateleng*), possessive (*sa Gaborone*)

3. Implementation

Figure 1 below shows a block diagram of the proposed tagger. Individual words are first tagged using morphological and noun analyzers developed in Malema et al (2016 and 2018). Simple compound POS are then tagged using regular expression (RE) Python library from Python NLTK. Regular expressions for simple compound POS are used here. We developed regular expressions for adjectives, enumeratives and for basic forms of possessives, relatives and adverbs. In Malema (2017) a finite state approach was used to tag basic multi-word POS. In this study we used the Python NLTK regular expression library because it is faster and much easier to use. After identifying compound POS in a sentence, expansion rules are applied to each compound POS for possible expansion. These rules basically test whether the next word(s) could be part of the current POS.

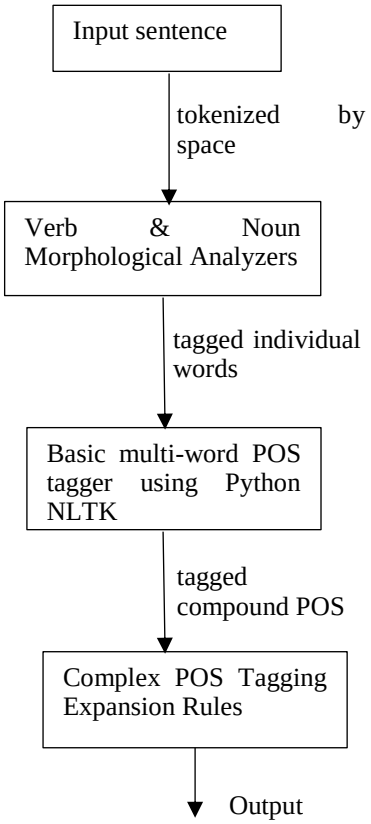

Figure 1: Block diagram of POS tagger

4. Performance Results

The proposed parts of speech tagger focused on complex direct relatives ending with a verb. The rule extensions developed are adding a noun, pronoun, qualificative, demonstrative or an adverb.
Examples:
ngwana yo o ratang (a child who likes …)
ngwana yo o ratang go lela (a child who likes crying)
ngwana yo o ratang dijo (a child who likes food)
ngwana yo o ratang dijo tse di sukiri (a child who likes sweet food)
ngwana yo o ratang dijo tsele (a child who likes that food)
And so forth.
As the examples show we focused on the basic structure of direct relatives which is *Relative concord + Verb-ng*. This structure could be extended in a variety of ways

Concord + Verb-ng + N
Concord + verb-ng + T
Concord + verb-ng + P
Concord + verb-ng + D
Concord + verb-ng + Q
Concord + verb-ng + L

Where N is noun, *T* is a qualificative, *P* is a pronoun, *D* is a demonstrative, *Q* is a quantitative and *L* is an adverb.
The prototype tagger was given a 300 sentence text file from the Botswana Daily News (2019) and Mmegi (2019). The text file contains 123 relatives, 37 of which are of the basic form and the rest are more complex. The proposed tagger identified all the 123 basic relatives and successfully extended 64 of the complex relatives resulting with a success rate of 74%. The two main factors that lead to the failure of the tagger are:

Unexhausted Relative forms:
We noted that there are other forms that we have not included in this structure. For example, we noted that there are forms in which the verb is followed by '*ke*' and '*le*' which are not in our rules. Examples:
yo o salang le ngwana (the one who is baby sitting)
yo o rutwang ke mmaagwe (the one taught by his/her mother)
Failure of basic word tagging:
In some cases the morphological analyzer failed to tag verbs, nouns and adverbs(single word) properly which affected the regular expression tagger and the expansion rule application. Also in some cases nouns were not put in their correct classes. The concord(s) of a qualificative modifying a particular noun has to match with its noun class.

5. Conclusions

In this paper we presented a rule based approach to identifying Setswana complex parts of speech. The idea is to implement the recursive structure of complex parts of

speech. The recursive structure is expressed in the form of rules which are based on simple verb and noun phrase structures. A prototype tagger was developed with the help of Python NLTK regular expressions. Preliminary results show that the proposed technique works well. However, for it to be effective, all the rules and structures of complex POS must be documented. In this study we did not exhaust all relative structures. We plan to develop the idea further by developing more rules and include other parts of speech.

6. Bibliographical References

Botswana Daily News (online), www.dailynews.gov.bw

Brants T (2000). A statistical part of speech tagger. *PANCL'00 Proceedings of the sixth conference on applied natural language processing Association for Computational Linguistics*

Brill E (1992). A simple rule based part of speech tagger. In *Proceedings of the third conference on Applied Natural Language processing, ACL*, Trento, Italy

Brill E (1995). Transformation Based Error-Driven Learning and Natural language Processing: A case study in Part of Speech Tagging. *Computational Linguistics*

Charniak E (1997).Statistical techniques for Natural Language parsin. *AI Magazine*, 18(4), pp.33-44

Cole, D.T. (1955). An Introduction to Tswana grammar. Longmans and Green, Cape Town.

Faaß G, Heid U, Taljard E & Prinsloo D (2009). Part-of_Speech tagging of Northern Sotho: Disambiguating polysemous function words",*Proceedings of the EACL, 2009 Workshop on Language Technologies for African Languages – Aflat 2009*, pages 38—45, Athens Greece, 31 March 2009

Lombard, D.P (1985). Introduction to the Grammar of Northern Sotho. J.L. van Schaik, Pretoria, South Africa, 1985.

Louwrens, L. J.(1991). Aspects of the Northern Sotho Grammar. Via Afrika, Pretoria, South Africa.

Mmegi Publishing News paper (online: www.mmegi.co.bw

Malema, G, Okgetheng, B and Motlhanka, M. (2017) Setswana Part of Speech Tagging, International Journal of Natural Language Computing (IJNLC), Vol.6, No.6, pp. 15 – 20, December 2017

Malema, G. Motlogelwa, N. Okgetheng, B. Mogotlhwane O. (2016). Setswana Verb Analyzer and Generator. International Journal of Computational Linguistics (IJCL), Vol 7, issue 1, 2016.

Malema, G. Motlhanka, M. Okgetheng, O and Motlogelwa, N. (2018). Setswana Noun Analyzer and Generator. International Journal of Computational Linguistics (IJCL), Volume (9), Issue (2) pp 32—40, 2018

Mogapi, K.(1998). Thuto Puo ya Setswana, Longman Botswana, 184, ISBN:0582 61903 3.

Pretorius L, Viljoen B, Pretorius R and Berg A.(2009). A finite state approach to Setswana verb morphology, *International Workshop on finite state methods and natural Language Processing FSMNLP 2009: Finite State Methods and Natural language Processing*, pp. 131 – 138

Taljard, E. & Bosch, S. E. (2006). A comparison of approaches towards word class tagging: Disjunctively versus conjunctively written Bantu languages. Nordic Journal of African Studies, 15(4), 428–442.