
Proceedings Paper:
Onyenwe, I.E. and Hepple, M. orcid.org/0000-0003-1488-257X (2016) Predicting Morphologically-Complex Unknown Words in Igbo. In: Sojka, P., Horák, A., Kopeček, I. and Pala, K., (eds.) Text, Speech, and Dialogue. Text, Speech, and Dialogue (TSD 2016), 12 - 16 Sep 2016, Brno, CzechRepublic. Lecture Notes in Computer Science, 9924 . Springer International Publishing , pp. 206-214. ISBN 978-3-319-45510-5

https://doi.org/10.1007/978-3-319-45510-5_24

Predicting Morphologically-Complex Unknown Words in Igbo

Ikechukwu E. Onyenwe and Mark Hepple

NLP Group, Computer Science Department, University of Sheffield, UK
{i.onyenwe,m.hepple}@sheffield.ac.uk

Abstract. The effective handling of previously unseen words is an important factor in the performance of part-of-speech taggers. Some trainable POS taggers use suffix (sometimes prefix) strings as cues in handling unknown words (in effect serving as a proxy for actual linguistic affixes). In the context of creating a tagger for the African language Igbo, we compare the performance of some existing taggers, implementing such an approach, to a novel method for handling morphologically complex unknown words, based on morphological reconstruction (i.e. a linguistically-informed segmentation into root and affixes). The novel method outperforms these other systems by several percentage points, achieving accuracies of around 92% on morphologically-complex unknown words.

Key words: Morphology, morphological reconstruction, Igbo, unknown words prediction, part-of-Speech tagging.

1 Introduction

The handling of unknown words is an important task in NLP, which can be assisted by morphological analysis, i.e. decomposing inflected words into their stem and associated affixes. In this paper, we address the handling of unknown words in POS tagging for Igbo, an *agglutinative* African language. We present a novel method for handling morphologically-complex unknown words of Igbo, based on morphological reconstruction (i.e. a linguistically-informed segmentation into root and affixes), and show that it outperforms standard methods using arbitrary suffix strings as cues.

In the rest of the paper, we first note prior work on unknown word handling in POS tagging, and consider the suitability of these methods to Igbo, as an agglutinative language. We then present some experiments using morphological reconstruction in unknown word handling for Igbo, and discuss our results.

2 Related Literature

Previous work on POS tagging unknown words has used features such as prefix and suffix strings of the word, spelling cues like capitalization, and the word/tag values of neighbouring words [9, 11, 4, 1]. The HMM method of Kupiec [5] assigns

probabilities and state transformations to a set of suffixes of unknown words. Samuelsson [10] used starting/ending n-length letter sequences as predictive features of unknown words. Brants [1] showed that word endings can predict POS, e.g. $-able$ is likely to be adjective in English. Toutanova *et al.* [11] uses variables of length up to n for extracting word *endings*, such that $n = 4$ for *negotiable* generates feature list `[e,le,ble,able]`. These methods have worked well in languages like English and German whose derivational and inflectional affixes reveal much about the grammatical classes of words in question.

3 Problem Description

Igbo is an agglutinative language, with many frequent suffixes and prefixes [3]. A single stem can yield many word-forms by addition of affixes, that extend its original meaning. Suffixes have different grammatical classes, and may concatenate with a stem in variable order, as in e.g.: *abịakwa "a-bịa-kwa", bịakwaghị "bịa-kwa-ghị", bịaghịkwa "bịa-ghị-kwa", bịaghachiri "bịa-gha-chi-ri", bịachighara "bịa-chi-gha-ra", bịaghachirịrị "bịa-gha-chi-rịrị"*, etc. Methods to automatically identify suffix-string cues (e.g. for use in POS tagging), based on extracting the last n letters of words, seem likely to be challenged by such complexity, e.g. that *bịaghachirịrị* "must come back" has 3 suffixes of length 3 or 4, to a total length of 10, which may elsewhere appear in a different order.

The Igbo POS tagset of Onyenwe *et al.* [7] uses "_XS" to indicate *extensional suffixes*, e.g. tag "VSI_XS" applies to a word that is VSI (verb simple) and includes $\geq$1 extensional suffix. In our experiments, using 10-fold cross validation and deriving the lexicon from the training data, we find that the majority of unknown words encountered arise due to agglutination (see Table 1).

4 Experiments

Our experiments compare methods using automatically-identified n-character suffix (and prefix) strings to methods based on (linguistically-informed) morphological reconstruction, in regard to their performance for handling morphological-complex Igbo words previously unseen in the training data during POS tagging.

4.1 Experimental Data

There are two sets of corpus data used in this research, the selected books from new testament Bible[1] that represents Igbo tagged religious texts (IgbTNT) and novel[2] for modern Igbo tagged texts genre (IgbTMT). The corpus data and the tagset used were developed in [8] and [7].

[1] Obtained from `jw.org`.
[2] "Mmadụ Ka A Na-Arịa" written in 2013.

4.2 Experimental Tools

POS Taggers We chose POS tagging tools that generally perform well, and have parameters to control word feature extraction for unknown word handling: Stanford Log-linear POS Tagger (SLLT) [11], Trigrams'n'Tagger (TnT) [1], HunPOS [4] (a reimplementation of TnT), and FnTBL [6] which uses the *transformation-based learning* (TBL) method of [2], adapted for speed. TBL starts with an initial state (where known words are assigned their most common tag, or a default) and applies *tranformation rules*, to correct errors in the initial state based on context. The training process compares the initial state tags to the true tags of the training data, and iteratively acquires an list of rules correcting errors in the initial state, until it sufficiently resembles the truth.

Morphological Reconstruction We used morphological reconstruction to segment morphologically-complex words into stems and affixes, so that patterns can be learnt over these sequences, which are used to predict the tags of unknown words. Items in these sequences classified as stem (ROOT), prefix (PRE) and suffix (SUF), i.e. ignoring finer distinctions of their grammatical function. For example, the word *enwechaghị* tagged "VPP_XS" in the IgbTC will have the form "e/PRE nwe/ROOT cha/SUF ghị/SUF" after morphological reconstruction. The idea is to use these morphological clues to predict the tag "VPP_XS", should the word appear as an unknown word.

For an inflected word w, morphological reconstruction involves extracting the stem cv and all n possible affixes attached to it. An Igbo stem is a formation of cv, starting with a consonant c and ending with a vowel v [3], where c could be a single letter or a digraph. Digraphs are two character strings pronounced as one sound and are non split (e.g."gh","ch"). We used a list of suffixes from [3] as a dictionary to search for valid morphological forms.

4.3 Experimental Setup

In our experiments, unknown words arise due to our use of 10-fold cross validation, i.e. the unknown words of any test fold are the words that were not present anywhere in the corresponding training set (i.e. the other 9 folds). Table 1 shows the unknown word ratios for our different data sets (listed under experiment1).

Table 1. Average sizes of train, test, and unknown words ratio of the experimental corpus data. Train2 and test2 are morphologically inflected data.

Corpus	Data for experiment1			Data for experiment2	
	Train	Test	Unknown Ratio	Train2	Test2
IgbTNT	35938	3993	3.18%	4120	088
IgbTMT	35965	3996	4.90%	4855	134
IgbTC	71902	7989	3.39%	8975	222

Our experiments compare the effectiveness of different methods for tagging such unknown words, and specifically the inflected ones. In our first experiment (ex-

periment1), we apply standard taggers to the data, and score their performance on the inflected unknown words. Our second experiment (experiment2) handles these same unknown words via morphological reconstruction. For this, we extract only the inflected unknown words from the data of experiment1, giving rise (under 10-fold cross validation) to the data set sizes listed under experiment2 of Table 1. (Note that these numbers might seem to be less than is implied by the "Unknown Ratio" column, as only the *inflected* unknown words are extracted, which correspond to around 70% of all unknown words.)

4.4 Experiment 1: Using Original Word-Forms

HunPOS, TnT and SLLT taggers were used because they have robust methods for extracting last/first letters of words for use as cues in handling unknown words. We chose n=5 and n=1 for extracting last and first letters of a word because the longest suffixes and prefixes in Igbo so far are of these lengths, and the taggers performed well at these settings. These systems also use the context of neighbouring words/tags to help in handling the unknown words. Table 2 shows the performance of these systems for the correct tagging of *only* the inflected unknown words (listed under experiment1).

Table 2. Average statistics and accuracy scores on the inflected tokens based on different approaches.

		1st Experiment			2nd Experiment				
Corpus	Size	HunPOS	TnT	SLLT	Taggers	PRE+SUF	PRE+SUF +rV	All	All(-PRE)
IgbTNT	88	70.73%	73.94%	83.77%	FnTBL	78.03%	82.81%	90.44%	82.78%
					SLLT2	66.31%	67.11%	66.53%	70.87%
IgbTMT	134	67.17%	70.37%	86.48%	FnTBL	78.96%	86.03%	91.99%	85.95%
					SLLT2	74.45%	75.27%	76.01%	77.15%
IgbTC	193	70.28%	73.16%	84.67%	FnTBL	83.75%	86.23%	88.46%	83.27%
					SLLT2	76.41%	77.62%	76.09%	76.54%

4.5 Experiment 2: Using Morphologically Reconstructed Forms

Our morphology segmentation module was used to perform morphological reconstruction of the data listed under experiment2 of Table 01. In the representation produced, the *correct* tag of an unknown word is marked on its *stem* within the stem/affix sequence. For example, *abịakwara* has tag VPP_XS, and so, after reconstruction, would be represented as "a/PRE bịa/VPP_XS kwa/SUF ra/SUF".

Four variants of the method were used, differing mostly in the extent to which the grammatical function of affixes were distinguished. In *Pattern1*, all affixes were classed as only either SUF (suffix) or PRE (prefix). In *Pattern2*, an "rV" tag was used for past tense suffixes.[3] In *Pattern3*, more morph-tags for suffixes

[3] Here, "rV" means letter *r* and any vowel (*a,e,u,o,i,ị,ọ,ụ*) attached to a word in Igbo like "bịara" *came*, "kọrọ" *told*, "riri" *ate*, "nwuru" *shone*, etc. It is a past tense marker if attached to active verb or indicate stative/passive meaning if attached to a stative verb [3]. Therefore, it is an important cue in predicting past tense verbs or verbs having applicative meaning "APP".

were added to indicate grammatical functions (see Table 4 for a list of the morph-tags). In *Pattern4*, prefix and stem were collapsed to form one part (e.g. changing "a/PRE bịa/VSI_XS kwa/LSUF" to "abịa/VSI_XS kwa/LSUF"), eliminating the "PRE" tag. Morph-tags serve as important clues for disambiguation.

Table 3. Some samples of morphological reconstructed words into stems and affixes.

Word form	FnTBL2 Initial State	FnTBL2 Truth State
	Pattern1 PRE+SUF	
nwukwasị	nwu/ROOT kwasị/SUF	nwu/VSI_XS kwasị/SUF
nwukwara	nwu/ROOT kwa/SUF ra/SUF	nwu/VrV_XS kwa/SUF ra/SUF
nwukwasịrị	nwu/ROOT kwasị/SUF rị/SUF	nwu/VrV_XS kwasị/SUF rị/SUF
ịnọdonwu	ị/PRE nọ/ROOT do/SUF nwu/SUF	ị/PRE nọ/VIF_XS do/SUF nwu/SUF
abịakwara	a/PRE bịa/ROOT kwa/SUF ra/SUF	a/PRE bịa/VPP_XS kwa/SUF ra/SUF
nụrụkwanụ	nụ/ROOT rụ/SUF kwa/SUF nụ/SUF	nụ/VSI_XS rụ/SUF kwa/SUF nụ/SUF
enwechaghị	e/PRE nwe/ROOT cha/SUF ghị/SUF	e/PRE nwe/VSI_XS cha/SUF ghị/SUF
	Pattern2 added "rV" to pattern1 and Pattern3 added all Morpho-tags	
nwukwasị	nwu/ROOT kwasị/LSUF	nwu/VSI_XS kwasị/LSUF
nwukwara	nwu/ROOT kwa/rSUF ra/rV	nwu/VrV_XS kwa/rSUF ra/rV
nwukwasịrị	nwu/ROOT kwasị/rSUF rị/rV	nwu/VrV_XS kwasị/rSUF rị/rV
ịnọdonwu	ị/PRE nọ/ROOT do/iSUF nwu/iSUF	ị/PRE nọ/VIF_XS do/iSUF nwu/iSUF
abịakwara	a/PRE bịa/ROOT kwa/eSUF ra/APP	a/PRE bịa/VPP_XS kwa/eSUF ra/APP
nụrụkwanụ	nụ/ROOT rụ/xSUF kwa/xSUF nụ/LSUF	nụ/VSI_XS rụ/xSUF kwa/xSUF nụ/LSUF
enwechaghị	e/PRE nwe/ROOT cha/xSUF ghị/NEG	e/PRE nwe/VSI_XS cha/xSUF ghị/NEG

Table 4. Morph-tags and meanings

Tag/Marker	Meaning
APP	Applicative
NEG	Negative
INFL	Inflection for perfect tense
rV	Inflection for past tense
LSUF	Last suffix marker for morphologically-inflected simple verb
xSUF	suffix within morphologically-inflected simple verb
eSUF	Suffixes within morphologically-inflected participle
iSUF	Suffixes within morphologically-inflected infinitive
rSUF	Suffixes within morphologically-inflected past tense verb

We applied FnTBL and SLLT to the morphologically reconstructed data (here referring to the latter as SLLT2, to differentiate from its earlier use in experiment 1). Note that the reconstructed representations for individual words are presented in isolation, i.e. so the systems cannot exploit contextual information of neighbouring words/tags (in contrast to experiment 1). FnTBL was chosen due to its effective pattern induction method, and SLLT because it outperformed the other systems in experiment 1. SLLT2 was simply trained directly over the reconstructed data. For FnTBL, we intervene to specify a particular initial state for TBL, in which the stem is given the initial tag "ROOT". Hence, TBL should generate only rules that, based on the morphological context, replace a ROOT tag with a final tag, the latter being a POS tag for a complete inflected unknown word. Results are shown in Table 2 under experiment2.

5 Discussion

Table 5 illustrates how the root+affixes have served as important cues for predicting the tags of morphological-complex unknown words. "Initial Tag" column is the FnTBL initial state, "Transformation Process" is predicted tag after applying transformational rules (adjacent to the stems are rules indexes that fired) and "Final Tag" is the FnTBL predicted tags returned as the tags for morphologically-complex unknown words. In Example 1 of Table 5, the word "begorochaa" refers to "perching activity of a group of birds", and is an inflected simple verb (VSI_XS) with "be" as the stem. Two transformational rules fired to transform its initial tag "ROOT" to the final tag "VSI_XS". The first change is made by Rule 0, which is a generic rule that changes ROOT tag to VrV (past tense verb) tag, provided there is a suffix within the [+1,+2] window. This rule is ordered first in the rule list, as it has the highest correction score over the training data. Rule 2 applies next, changing VrV to "VSI_XS" because xSUF and LSUF occur with inflected simple verbs. In other examples, Rule 2 changes VrV to VrV_XS (inflected past tense verbs) because of rSUF that occur in past tense verbs, Rules 3 and 4 change VrV and VSI_XS to VPP_XS (inflected participle) tag whenever the previous tag after stem is PRE, Rule 5 changes VPP_XS to VPERF (perfect tense verbs) due to presence of INFL, Rule 6 changes VPP_XS to VIF_XS (inflected infinitive verb) due to *ị* prefix, and Rule 36 changes VPERF to VPERF_XS (inflected perfect tense) due to xSUF.

Table 5. Examples of transformational rules generated by FnTBL. The numbers are the rule identity numbers that fired.

Initial Tag	Transformation Process	Final Tag
Example1		
be ROOT VSI_XS	be VSI_XS VSI_XS — 0 1	begorochaa/VSI_XS
go xSUF xSUF	go xSUF xSUF	
ro APP APP	ro APP APP	
chaa LSUF LSUF	chaa LSUF LSUF	
Example2		
kpọ ROOT VrV_XS	kpọ VrV_XS VrV_XS — 0 2	kpọchibidoro/VrV_XS
chi rSUF rSUF	chi rSUF rSUF	
bi rSUF rSUF	bi rSUF rSUF	
do rSUF rSUF	do rSUF rSUF	
ro rV rV	ro rV rV	
Example3		
e PRE PRE	e PRE PRE	ekpochapụ/VPP_XS
kpo ROOT VPP_XS	kpo VPP_XS VPP_XS — 0 3	
cha eSUF eSUF	cha eSUF eSUF	
pụ eSUF eSUF	eSUF eSUF	
Example4		
ị PRE PRE	ị PRE PRE	ịkpọcha/VIF_XS
kpọ ROOT VIF_XS	kp VIF_XS VIF_XS — 0 1 4 6	
cha LSUF LSUF	cha LSUF LSUF	
Example5		
e PRE PRE	e PRE PRE	echekwala/VPERF_XS
che ROOT VPERF_XS	che VPERF_XS VPERF_XS — 0 3 5 36	
kwa xSUF xSUF	kwa xSUF xSUF	
la INFL INFL	la INFL INFL	

The accuracy scores of both experiments are shown in 1st and 2nd experiment columns of Table 2. "PRE+SUF" column is for *Pattern1* variation, the accuracy scores are substantive, FnTBL did better than SLLT2 in all cases and performed better than other taggers in experiment 1 except SLLT. Column "PRE+SUF+rV" shows *Pattern2* variation, SLLT2 and FnTBL performances generally improve and FnTBL scored better than majority in 1st experiment. "All" column is for *Pattern3*, here is to test the prospect of paradigmatic tagging where meaningful tags for affixes are added to indicate their grammatical functions. This gave best scores of 90.44%, 91.99% and 88.46% for FnTBL and these scores are several points better than scores achieved by the taggers used in the 1st experiment (see Table 2).

Finally, column "All(-PRE)" for *Pattern4* is to verify the strength of prefix as unknown word predictive feature considering it is only one character length. Comparing columns "All(-PRE)" and "All", shows that there are lost in accuracies of column "All" for FnTBL (e.g. about 9.0 in IgbTNT). This is contrary to English where addition of prefix as feature caused negative effect on the accuracy of unknown words [11]. Surprisingly, SLLT2 increased in its accuracy against decrease in FnTBL scores. But an experiment on IgbTMT using SLLT tagger's technique for handling unknown words shows that using only suffix features gave accuracy of 77.26% and addition of prefix features improved the accuracy on the morphologically-complex words by 9.22%. The reason for SLLT2's accuracy increment can be explained in regard with "PRE" ambiguity. "PRE" tag is used to indicate prefix whether it is "ị/i" for infinitive or "a/e" for participle and simple verbs, therefore, collapsing it with the stem removes this ambiguity. Statistical taggers will require large data size to properly disambiguate this case.

6 Conclusion

We have shown that use of actual linguistically-informed segmentation into stems and associated affixes are good for predicting unknown inflected words in Igbo. Through morphological reconstruction, inflected words are represented in machine learnable pattern that exploits morphological characteristics during tagging process for handling unknown words. The performance of FnTBL that inductively learns linguistic patterns reveals that our method is better than methods that automatically identify suffix-string cues (e.g. for use in POS tagging), based on extracting the last n letters of words to serve as proxy for actual linguistic affixes. The standard method using arbitrary suffix strings as cues is challenged by complexity associated with morphologically-complex unknown words of the language. In Igbo language, a single root can produce as many possible word-forms as possible through the use of affixes of varying lengths ranging from 1 to 5, which may concatenate with a stem in variable orders.

In the future work, it is important to perform full morphological analysis on Igbo. This experiment excludes some inflected classes (like nouns) as it will lead to full morphological analysis which is beyond the research scope. Also, morphological analysis on the compound verbs and exploiting n neighbouring words

information are ignored. These lapses will hide some important information required for NLP task. Of course, this is pointing towards building a large-scale computational morphologies for Igbo.

Acknowledgments

We acknowledge the financial support of Tertiary Education Trust Fund Nigeria and Nnamdi Azikiwe University (NAU) Nigeria. Many thanks to Dr. Uchechukwu Chinedu of linguistic department, NAU for his very helpful discussion.

References

1. Brants, T. TnT: A Statistical Part-of-speech Tagger. In: Proceedings of the Sixth Conference on Applied Natural Language Processing. 224–231 (2000)
2. Brill, E. Transformation-based Error-driven Learning and Natural Language Processing: A Case Study in Part-of-speech Tagging. Comput. Linguist. 21, 543–565. MIT Press, Cambridge, MA, USA (1995)
3. Emenanjo, N. E. Elements of Modern Igbo Grammar: A Descriptive Approach. Ibadan Oxford University Press (1978)
4. Halácsy, P., Kornai, A., Oravecz, C. HunPos: an open source trigram tagger. In: Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions. 209–212 (2007)
5. Kupiec, J. Robust Part-of-Speech Tagging Using a Hidden Markov Model. In: Academic Press Limited (1992)
6. Ngai, G., Florian, R. Transformation-based learning in the fast lane. In: Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies. 1–8 (2001)
7. Onyenwe, I. E., Uchechukwu, C., Hepple, M. Part-of-speech Tagset and Corpus Development for Igbo, an African. LAW VIII. 93. (2014)
8. Onyenwe, I. E., Hepple, M., Uchechukwu, C., Ezeani, I. Use of Transformation-Based Learning in Annotation Pipeline of Igbo, an African Language. In: Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects. 24 (2015)
9. Ratnaparkhi, A., others. A maximum entropy model for part-of-speech tagging. In: Proceedings of the conference on empirical methods in natural language processing. 1, 133–142 (1996)
10. Samuelsson, C. Morphological tagging based entirely on Bayesian inference. In: 9th Nordic conference on computational linguistics. (2013)
11. Toutanova, K., Klein, D., Manning, C. D., Singer, Y. Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. 1, 173–180 (2003)