Améliorer la précision d'annotation d'un corpus Igbo par reconstruction morphologique et l'apprentissage basé sur la transformation

Ikechukwu Onyenwe[1] Mark Hepple[1] Uchechukwu Chinedu[2]
(1) Natural Language Processing Lab,
Computer Science Department,
University of Sheffield, United Kingdom
(2) Linguistic Department
Nnamdi Azikiwe University, Nigeria.
i.onyenwe@sheffield.ac.uk, m.hepple@sheffield.ac.uk, neduchi@yahoo.com

RÉSUMÉ

Cet article présente une méthode employée pour améliorer l'exactitude des tags syntaxiques d'un corpus Igbo (une langue Africaine agglutinante), en se concentrant sur les formes fléchies. Dans un premier temps, une reconstruction morphologique est appliquée pour identifier ces formes fléchies et les segmenter (segmentation linguistique sous forme de racines et d'affixes). Celle-ci est ensuite utilisée conjointement avec l'algorithme d'apprentissage TBL (Transformation-Based Learning) pour identifier les mots mal labellisés dans le corpus et proposer une étiquette de remplacement, de sorte qu'un expert humain pourrait ensuite accepter ou rejeter ces changements. Pour évaluer l'impact de ce procédé sur le corpus, nous l'avons utilisé pour entraîner un tagueur syntaxique. Durant la labellisation de l'Igbo, la majorité des mots inconnus (c'est-à-dire les mots non présents dans les données d'entraînement) apparaît en raison de la flexion. Nous avons observé grâce à notre approche une amélioration de la précision de 77,77% à 83,13% pour la labellisation syntaxique des mots inconnus, et de 58,01% à 86,81% sur les formes fléchies inconnues.

ABSTRACT

Improving Accuracy of Igbo Corpus Annotation Using Morphological Reconstruction and Transformation-Based Learning.

This paper describes a method that has been used to improve the correctness of a part-of-speech (POS) tagged corpus of Igbo (an agglutinative African language), focussing on inflected words. First, morphological reconstruction (a linguistically-informed segmentation into roots and affixes) was applied to identify inflected words and segment them. Then, this information was used together with transformation-based learning (TBL, a machine learning algorithm) to flag words that were candidates for having been incorrectly tagged in the corpus, and to suggest a replacement tag, so that a human expert could accept or reject these changes. To assess the impact of this process on the corpus, we used it to train a POS tagger. During Igbo tagging, the majority of unknown words (i.e. words not seen in the training data) arise due to inflection. We found that tagging accuracy for unknown words increased from 77.77% to 83.43%, and for inflected unknown words from 58.01% to 86.81%.

ABSTRACT

Igbo : Nkwarite Ọdị Mma Igbo Corpus Site N'iji Morphological Reconstruction Na Transformation-Based Learning.

Edemede a na-akọwa usoro nke e ji kwarite ndịwanye mma nke nkejiasụsụ-okwu akpadoro na corpus Igbo (asụsụ agglutinative nke Africa), tụmadị n'okwu ndị nwere mọfịm. Mbụ, e ji morphological reconstruction (bụ maka ikewapụta mgbọrọgwụ-okwu na irihiri-okwu (mọfịm) dị n'ime okwu nwere ha) wee mee ka amata okwu ndị nwere irihiri-okwu. Mgbe nke a gasịrị, amamihe dị n'ime nkewapụta a na Transformation-Based Learning (TBL, igwe mmụta algọridim ọmụmụ) jikọrọ aka ọnụ wee rụba aka ebe niile okwu ndị nwere mọfịm n'ime ha ma ha enweghị ezigbo mkpado nkejiasụsụ-okwu nke n'egosipụta na ha nwere mọfịm, na na-atụkwa aro nnọchi mkpado nkejiasụsụ-okwu, nke mere na mmadụ ọkachamara nwere ike ịnabata ma ọ bụ jụ mgbanwe ndị a. Ịmata mmetụta nke usoro a na corpus a, anyị ejiri ya zụa POS tagger. N'oge ọzụzụ POS tagger, ọtụtụ okwu ndị ana amaghị (okwu ndị POS tagger na-ahụghị oge ọzụzụ) bụ maka mọfịm dị n'ime ha. Anyị chọpụtara na mkpado ziri ezi maka okwu ndị niile ana amaghị dịwanyere mma site na 77.77% rue 83.43% na okwu ndị ana amaghị nwere mọfịm dịwanyekwara mma site na 58.01% rue 86.81%.

MOTS-CLÉS : Corpus, partie du discours, l'apprentissage basé sur la transformation, l'apprentissage de la machine, Igbo, morpho Reconstruction logique, Morphologie ou affixes, mots inconnus.

KEYWORDS: Corpus, Part-of-speech, Transformation-based learning, Machine learning, Igbo, Morphological Reconstruction, Morphology or Affixes, Unknown words.

1 Introduction

The availability of high quality of part-of-speech (POS) annotated corpora is crucial to the development of automatic POS taggers. POS taggers choose the tag to assign in ambiguous cases (i.e. words having more than one possible tag) by using *context*. For example, Hidden Markov Model (HMM) taggers typically use information of the preceding two words, and their tags, in this way. Consequently, errors in the POS tag assigned to words in a corpus present a threat to creating effective taggers, as they give rise to 'false contexts' about which the tagger will learn, and which stand in place of a 'true context', which could have provided valuable evidence in training the tagger (Pavel & Karel, 2002).

Errors in annotated corpora arise during the annotation process, and the error-rate will be partly a function of the way in which a corpus was created. There are different options for creating a POS annotated corpus : it could be annotated entirely manually, or entirely automatically, or a mixed approach could be used (combining manual and automatic methods). Manual annotation is in general expected to give a lower error-rate, but is very labour-intensive and costly, limiting the size of the corpus that can be produced. Automatic annotation can be applied to produce much greater volumes of annotated material than could ever be manually annotated, but the error-rate will be higher than for manual annotation. In semi-automatic annotation, human input might be given at different stages of the overall process, e.g. in creating some initially manually tagged material to train a tagger which is applied to further materials, and in manually checking the output of that tagger, to ensure an acceptable error-rate. Despite the best efforts of annotators during pre- and post-editing phases, tagged corpora will inevitably still contain errors. There is therefore a value in developing methods to automatically detect likely errors in a corpus, and to suggest possible corrections, for inspection by a human expert. This will be much more efficient, and therefore more feasible, than asking an expert annotator to search methodically for errors.

In this paper, we propose an automatic method for locating likely errors in the assignment of POS tags to morphologically-inflected words in Igbo. The method uses morphological reconstruction, a linguistically-informed segmentation into roots and affixes, together with transformation-based

learning (TBL), a machine learning algorithm due to Brill (1995). Igbo is an agglutinative language, in which morphological affixation produces many variant vocabulary items. Igbo affixes vary in length from 1 to 5 characters, and the attachment of multiple affixes, in variable order, to a single stem, can produce many different forms, which extend the stem's original meaning (see Table 1). We used morphological reconstruction as a means to identify and correct inflected words that have been incorrectly tagged in the Igbo tagged corpus (Onyenwe *et al.*, 2014, 2015). Representing words as sequences of stems and affixes allows patterns for predicting their POS to be automatically acquired.

Transformation-based learning (Brill, 1995) starts with correctly tagged text (*truth*) and an *initial state* tagging of the same material. For POS tagging, the initial state typically assigns each word its most common tag (or a default tag for unknown words). The training process iteratively acquires an ordered list of rules that correct errors found in the initial state, until this resembles the truth to an acceptable degree. In our experiments, we use the FnTBL ('TBL in the fast lane') software of Ngai & Florian (2001), which is a much faster reimplementation of Brill's original method. Morphological reconstruction segments words into units that can be exploited by TBL's linguistic pattern detection.

Word-form	Stem and Affixes	Meaning
ri	ri	eat
iri	i+ri	to eat
ga-eri	ga-+e+ri	will eat (auxiliary verb hyphnated to participle)
ga-ericha	ga-+e+ri+cha	will eat completely
ga-erichaṛịṛị	ga-+e+ri+cha+rịṛị	will must eat completely
ga-erikwa	ga-+e+ri+kwa	will eat also
richaṛịṛị	ri+cha+rịṛị	must eat completely
richakwa	ri+cha+kwa	eat completely also
richara	ri+cha+ra	ate completely
richakwara	ri+cha+kwa+ra	ate completely also

TABLE 1 – Illustrating word formation in Igbo using morphology

2 Igbo Language

Igbo is one of the major languages if eastern Nigeria, with around 32 million native speakers. [1] It is classified as a Niger-Congo language of the Kwa sub-group. [2] It has 28 consonants and 8 vowels, and uses the Ọnwụ Committee orthography. [3] See Onyenwe *et al.* (2014, 2015) for a detailed description.

3 Related Work

Morphological analysis has been usefully exploited elsewhere in natural language processing. Thede & Harper (1997), working on the TIMIT corpus, investigated whether morphological information could assist in handling unknown words in the context of syntactic parsing, and found their parser's performance to be greatly enhanced. Their morphological analysis method used only a knowledge of affixes in guessing the POS of unknown words, i.e. it did not use direct information about word stems.

1. en.wikipedia.org/wiki/Igbo_people [May, 2016]
2. www.igboguide.org/HT-igbogrammar.htm [May, 2016]
3. www.columbia.edu/itc/mealac/pritchett/00fwp/igbo/txt_onwu_1961.pdf [May, 2016]

Milne (1986) used morphological reconstruction to resolve ambiguity during parsing. Light (1996) used various knowledge sources to determine word meanings, including morphological cues.

Previous work on improving the correctness of annotated corpora has sought to make most efficient use of having a human expert in the loop, by using automated methods to identify errors and/or suggest corrections. Brill & Marcus (1992) used a semi-automatic approach for tagging an unfamiliar text and then applied learned rules to both correct errors and find where contextual information can repair tagging mistakes with limited help from a native speaker. Taljard *et al.* (2008) and Heid *et al.* (2006) used a lexicon containing 7000 known words and their annotations, and a noun and verb guesser, to pre-tag 40000 tokens of Northern Sotho's texts. The output was reviewed manually and correct guesses added to the lexicon, so that the lexicon continuously grows. Loftsson (2009) and Helgadóttir *et al.* (2012) applied trained POS taggers, singly and combined, respectively, and compared their outputs to gold standard tags, so that differences could be marked as error candidates for verification. Leech *et al.* (1983) used three stages to perform the overall tagging process of the Lancaster-Oslo/Bergen (LOB) corpus. The raw corpus started with manual computer-aided pre-editing, where a human inspector prepares it for input for automatic tagging using Tagged Brown Corpus, then the output of the automatic tagging (tagged corpus) is subjected to manual computer-aided post-editing where a human inspector corrects any error made during automatic tagging. In our experiment, we applied an automatic method that learns rules from the morphologically reconstructed words in Igbo tagged corpus (IgbTC) and then applied these rules to find and propose tags for all morphologically-inflected words not tagged properly. All positions where these changes occurred are inspected and corrected by human annotator expert for quality assurance.

4 Current State of Igbo Corpus

Igbo corpus (IgbC) is about 1 million tokens in size. A 263856 token segment of it was initially tagged by group of Igbo linguists using the tagging scheme and corpus developed by Onyenwe *et al.* (2014). We shall refer this segment as the Igbo tagged corpus (IgbTC). Then using the by-products [4] from inter-annotation agreement [5] (IAA) exercise and FnTBL, the quality of IgbTC was improved (Onyenwe *et al.*, 2015). Also, the committee-of-taggers method of Loftsson (2009) and Helgadóttir *et al.* (2012) was adopted for further improvement of IgbTC. There are two phases involved in the improvement process that used IAA by-products and FnTBL (IAA+TBL) : one is where majority of the linguists are agreeing on a tag and the other is where one linguist is disagreeing against others. The justification of the later is that this one linguist might have better insight on the usage of a particular tag than others. The main objective here is to train FnTBL on the linguist-annotated texts (a subset of IgbTC) and used FnTBL learned rules to track positions in the IgbTC where IAA revised tags could be imported. In summary, in the first phase, we used voting technique on the linguists' annotated texts to get majority votes on a tag and FnTBL was applied on the outcome. From Table 2, the result shows that there are 25490 positions inspected on the IgbTC with 19921 effective changes. That means this process flagged 25490 positions in the IgbTC ($\approx$10% of IgbTC) for human expert to inspect instead of examining methodically the positions in the IgbTC where revised tags changes are to be reflected. And about 78% of these flagged positions were effectively changed, that is to say, 7.550% errors were eliminated from IgbTC. Hence, this process is effective error correcting method. In the second phase where one linguist is against the others, 26155 positions were inspected with an effective change of

4. They are POS annotated sub corpora by the linguists. Sub corpora because they are selected texts from untagged IgbTC.
5. IAA was conducted for the purpose of making the tagging scheme reliable, reproducible and viable.

5684 made in IgbTC. The rate of corrections (≈22%) in the second phase is substantially lower than first phase, but the number of corrections (2.154% errors were eliminated from IgbTC) justifies this as an effective process. Results are shown in Table 2.

When the FnTBL rule fired at a location, it proposed a specific revised tag change, then the human annotator expert can either accept the FnTBL proposed revised tag "Accepted Judgement", retain the existing tag at the location where the current tag in IgbTC is deemed correct "No-Change Required", or impose an alternative change according to his knowledge of revised tagset when neither FnTBL proposed tag or current tag in IgbTC are correct "Manual Change". The "Effective Change" column is calculated by adding columns "Accepted Judgement" and "Manual Change".

Name	Location Flagged	Accepted Judgement	No-Change Required	Manual Change	Effective Change	% Error Eliminated from the Corpus
IAA+TBL Phase1	25490	16612	5569	3309	19921	7.550
IAA+TBL Phase2	26155	3605	20471	2079	5684	2.154
COT	11810	6549	4165	1096	7645	2.897
Total	63455	26766	30205	6484	33250	12.601

TABLE 2 – Total statistics outcomes of the improvement methods.

For committee of taggers (COT) experiment, Stanford Log-linear Tagger (Toutanova *et al.*, 2003), MBT– A memory-based POS tagger-generator by (Daelemans *et al.*, 1996), and FnTBL–Transformation-based learning in the fast lane by (Ngai & Florian, 2001) were used. The taggers flagged tags in 11810 positions in IgbTC where they have disagreement. After inspection, an effective change of 7645 was made to improve IgbTC.

The entire improvement processes resulted in inspecting 24.05% of IgbTC with 12.601% effective change made and accuracy increased from 88% (initial state of IgbTC) to 96% (current state of IgbTC) obtained by training and testing FnTBL tagger on IgbTC sets on 10-fold cross validation over the corpus size. Part of these improvement processes have been reported in (Onyenwe *et al.*, 2015).

5 Experiment

5.1 Morphological Segmentation

We designed a module for segmenting morphemes and stems of morphologically-inflected words found in IgbTC such that their stems and affixes are classified as stem (ROOT), prefix (PRE) and suffix (SUF) tags irrespective of their grammatical functions, this will generate a tagset of {PRE ROOT $SUF_{i...n}$} for any given morphologically-inflected word. Prefix in Igbo is only a single character long. For example, this word *enwechaghị* tagged "VPP_XS" in the IgbTC will have the form "e/PRE nwe/ROOT cha/SUF ghị/SUF" after morphological reconstruction. The plan here is to use these morphological clues to predict the correct tags for the morphologically-inflected words.

The approach is, for any given word w, the stem cv is extracted and all n possible morphological parts attached to cv are generated. Stem in Igbo is a formation of cv that starts with a consonant c and ends with a vowel v (Emenanjo, 1978), where c could be a single letter or double in the case of digraph. Digraphs are two character strings pronounced as one sound and are non split (examples "gh", "ch", "kw", "gb", "gw", "nw", "ny", "sh", "kp"). We used a list of suffixes from (Emenanjo,

1978) as a dictionary to search for valid morphological forms. To test how robust this system is, we avoided using any tag information from IgbTC for tracking of morphologically-inflected words. Therefore, for any given word, if there is n valid morphological part(s) attached to its cv, then the word will be reconstructed ("e/PRE nwe/ROOT cha/SUF ghị/SUF"). Otherwise, that word is not morphologically-inflected.

This is not a full scale computational morphology in Igbo, we only focused on morphologically-inflected words that are verbs since they constitute the majority of words in the morphologically-inflected class. We avoided full scale morphological analysis at this stage because of time constraints. Morphologically-inflected verbs are extracted from the corpus (IgbTC) by having in the suffix list only morphemes that are associating with verbs. In case of verbs' nominalization to nouns, we used nominalizing prefixes (n,m,o,u,ọ,ụ) to track these instances not entering for reconstruction. Another important clue is the use of word-shape, verbs shape normally starts with *VCV, CV, CVV, VCVCV, CVCVCV* ("C" is consonant and "V" is vowel) and so on but cannot end with a C. For example, verbs "atụkwasịrị" and "banyekwa" have common word-shapes of "VCVCVCVCV" and "CVCVCV" for verbs but words "mpịakọta", "Kapanịọm" and "mgbaasị" have word-shapes "CCVVCVCV", "CCVVCV" and "CVCVCVVC" different from verbs.

How accurate is this system in tracking morphologically-inflected words that are verbs ? Igbo tagset is designed to have special tags given to morphologically-inflected words. We used this information to build lexicon of all morphologically-inflected words that are verbs and compared it with the output of morphological segmentation system. For example, there are 31383 morphologically-inflected verbs in the IgbTC, segmentation system extracted 35208 words from this corpus, and out of this number, 29817 (95.01%) are morphologically-inflected verbs and 5391 are not. The remaining 4.99% of morphologically-inflected verbs require more robust computational morphology to segment properly. The 5391 words wrongly identified as being inflected verbs have same word-shape with verbs. For example, "ọchịchịrị" is a noun with shape "VCVCVCV" same as verb "ekwusakwa" (VCVCVCV). And most of words found in this 5391 words are nouns (mainly common), we used list of noun class constructed from the corpus to eliminate them.

5.2 Improving the Accuracy of the Current State of Igbo Tagged Corpus

Igbo tagset is defined in two parts : α and α*_XS*, where α represent any non morphologically-inflected tag and *XS* is to indicate presence of any affix in a word that is morphologically-inflected (Onyenwe *et al.*, 2014). This experiment automatically find and correct those morphologically-inflected words that suppose to be tagged α*_XS* but are not in IgbTC. We used following tools : morphological segmentation discussed above, transformation-based learning on the fast lane (FnTBL) (Ngai & Florian, 2001) (a reimplementation of (Brill, 1995)) and Stanford Log-linear Tagger (SLLT) (Toutanova *et al.*, 2003). The output of morphological segmentation will benefit FnTBL's linguistic pattern detection. SLLT has a robust technique for handling unknown words. It uses variables up to n in extracting first/last letters of a word such that n = 4 for *negotiable* will generate extraction list [e,le,ble,able] to serve as proxy for linguistic affixes.

FnTBL was trained and tested on the outputs of morphological segmentation module (see outputs in Table 3). FnTBL's initial state takes the output where "ROOT" is given to all the verb stems while the associated affixes are given SUF (suffixes) and PRE (prefixes), and then ROOT will be replaced with the original verb's tag from IgbTC in the FnTBL's truth state. For example, in Table 3, the verb nwukwasị tagged "VSI_XS" in the IgbTC will have forms "nwu/ROOT kwasị/SUF" and

	Morphologically Reconstructed	
Word form	FnTBL Initial State	FnTBL Truth State
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
ịzụpụtara	ị/PRE zụ/ROOT pụ/SUF ta/SUF ra/SUF	ị/PRE zụ/VIF_XS pụ/SUF ta/SUF ra/SUF

TABLE 3 – Some samples of morphological-complex words morphologically reconstructed into stems and affixes to serve as FnTBL states.

"nwu/VSI_XS kwasị/SUF" for FnTBL's initial and truth states respectively.

Initial Tag	Transformation Process	Final Tag
Example1 nwu ROOT VrV kwasị SUF SUF kwa SUF SUF ra SUF SUF	 nwu VrV_XS VrV- 0,2 kwasị SUF SUF kwa SUF SUF ra SUF SUF	nwukwasịkwara/VrV_XS
Example2 zu ROOT VrV kwa SUF SUF a SUF SUF ra SUF SUF	 zu VrV_XS VrV- 0,2 kwa SUF SUF a SUF SUF ra SUF SUF	zukwaara/VrV_XS
Example3 fu ROOT NNC nahụ SUF SUF	 fu VSI_XS NNC - 0,1 nahụ SUF SUF	funahụ/VSI_XS

TABLE 4 – Some examples of FnTBL error correction process using morphological information. The numbers in the middle column are FnTBL rule's index number.

Observation on the FnTBL's output revealed that there are verbs in IgbTC that their tags are signifying that they are not morphologically-inflected but FnTBL is indicating otherwise by assigning them morphologically-inflected tags (α*_XS*). For example, "Transformation Process" column in Table 4, 0 is an FnTBL rule that changes ROOT to VrV (past tense verb), 2 is another FnTBL rule that changes VrV to VrV_XS if there is any SUF and last SUF that has rV form. rV means letter "r" and any vowel (a,e,i,ị,o,ọ,u,ụ) which is a past tense marker in Igbo (Ikegwxqnx, 2011). FnTBL rule 1 changes VrV to VSI_XS (morphologically-inflected simple verbs) if there is any SUF. We automatically verified this, for each instance flagged by FnTBL as being morphologically-inflected, if both tags (IgbTC tag and FnTBL suggested tag) have same α tag and there is any SUF, then we chose FnTBL suggested tag.

For instance, in Tables 4 and 5, nwukwasịkwara/VrV and zukwaara/VrV were tagged "VrV" in the corpus (IgbTC) but FnTBL suggested "VrV_XS". In both cases, VrV is the α tag and there is existence of affixes (SUF), therefore "VrV_XS" tag will be chosen. After FnTBL transformational process, all ROOT tags are changed to underlined tags in "Final Tag" column of Table 4. The examples with ROOT tag are FnTBL input (initial state) that came from morphological segmentation module (see Table 3). Another interesting example is "ịhapụrụ" where FnTBL suggested the right tag "VIF_XS" (morphologically-inflected infinitive verb) using the prefix "ị" information even though

IgbTC Before Error Correction	IgbTC After Error Correction
nwukwasịkwara/VrV	nwukwasịkwara/VrV_XS
pụkwaghị/VrV_XS	pụkwaghị/VSI_XS
bụrụkwa/VrV_XS	bụrụkwa/VSI_XS
laara/VrV	laara/VrV_XS
waara/VrV	waara/VrV_XS
zooro/VrV	zooro/VrV_XS
zukwaara/VrV	zukwaara/VrV_XS
kwughachikwa/VCO	kwughachikwa/VSI_XS
kwuluwo/VSI_XS	kwuluwo/VPERF
ịhapụrụ/VrV_XS	ịhapụrụ/VIF_XS
kwọọ/VSI	kwọọ/VSI_XS
gbawasịa/VrV	gbawasịa/VSI_XS
tọgbọbu/VrV	tọgbọbu/VSI_XS
funahụ/NNC	funahụ/VSI_XS
tachie/NNCV	tachie/VSI_XS

TABLE 5 – Sample of morphologically-inflected words corrected.

the last two letters usually indicates VrV_XS or VrV tag. The remainders (like kwughachikwa/VCO and kwughachikwa/VSI_XS, bụrụkwa/VrV_XS and bụrụkwa/VSI_XS in Table 5 where there are different α) were manually corrected. With this data improvement method, we corrected a total of 380 samples (all morphologically-inflected) in IgbTC. For quality assurance, all these positions were inspected by a human annotator expert.

For training and testing SLLT on IgbTC, IgbTC was sets into train and test data on a 10-fold cross validation over the corpus size. The unknown word ratio is the percentage of words previously unseen in the train data.

Table 6 shows the results when we applied SLLT on the IgbTC. Before and After Error Correction columns show accuracies before and after application of this improvement method on IgbTC. After the application of this error correction process, SLLT accuracy scores on IgbTC generally increased. The effect is very prominent in the accuracy of the unknown words (especially the inflected words (see Table 6)).

	IgbTC State Before Error Correction			IgbTC State After Error Correction		
Taggers	Overall Scores	Unknown Scores	Inflected Unknown Words Scores	Overall Scores	Unknown Scores	Inflected Unknown Words Scores
SLLT	98.05%	77.77%	58.01%	98.11%	83.43%	86.81%

TABLE 6 – Data improvement accuracy results using SLLT on the IgbTC

Observation from the impact of this experiment from Table 6 shows that the majority of the corrected tags belong to the unknown words class which are mostly morphologically-inflected words that are less frequent. The accuracy scores after this error correction method show that SLLT gained extra 0.06% for overall, 5.66% for unknown words and 28.8% for morphologically-inflected words that are unknown. The accuracy scores are not about experiment in handling unknown words, rather we are showing the level of effects of this error correction technique on the sides of unknown words (both those that morphologically-inflected) and overall words.

6 Conclusion

In this paper, we have shown how we used morphologically reconstructed words into stems and associated affixes together with TBL to transform wrongly tagged morphologically-inflected words to their correct tags in tagged Igbo corpus (IgbTC). Through morphological reconstruction, a more linguistically-informed segmentation into roots and affixes, morphologically-inflected words in IgbTC are represented in machine learnable pattern that FnTBL exploited to identify and suggest plausible tags for those tags assignment to the morphologically-inflected words that violated their true status. Human annotator expert inspected all the affected positions on IgbTC for quality assurance.

Further work is how best to maximize the use of morphological characteristics in Igbo to increase the accuracy of tagging unknown words. The accuracy of SLLT on the unknown words is lower than other languages (like English) it has been tested on. This is towards building a full scale computational morphology for Igbo.

Acknowledgement

We acknowledge the financial support of Tertiary Education Trust Fund (TETFund), Nigeria and Nnamdi Azikiwe University (NAU), Awka, Nigeria. Part of the improvement methods that gave the current state of tagged Igbo corpus (IgbTC) used in this experiment has been reported in Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects.

Références

BRILL E. (1995). Transformation-based error-driven learning and natural language processing : A case study in part-of-speech tagging. *Comput. Linguist.*, **21**(4), 543–565.

BRILL E. & MARCUS M. (1992). Tagging an unfamiliar text with minimal human supervision. In *Proceedings of the Fall Symposium on Probabilistic Approaches to Natural Language.*

DAELEMANS W., ZAVREL J., BERCK P. & GILLIS S. (1996). Mbt : A memory-based part of speech tagger-generator. *arXiv preprint cmp-lg/9607012*.

EMENANJO N. E. (1978). *Elements of Modern Igbo Grammar : A Descriptive Approach*. Ibadan Oxford University Press.

HEID U., TALJARD E., & PRINSLOO D. J. (2006). Grammar-based tools for the creation of tagging resources for an unresourced language : the case of northern sotho. In *5th Edition of International Conference on Language Resources and Evaluations*.

HELGADÓTTIR S., LOFTSSON H. & RÖGNVALDSSON E. (2012). Correcting errors in a new gold standard for tagging icelandic text. In *LREC'14 : 2944-2948*.

IKEGWXQNX C. N. (2011). Tense as an element of infl phrase in igbo. *Journal of Igbo Language and Linguistics (JILL)*, **3**, 112–121.

JURAFSKY D. & MARTIN J. H. (2014). *Part of Speech Tagging*. `https://web.stanford.edu/~jurafsky/slp3/9.pdf`.

LEECH G., GARSIDE R. & ATWELL E. S. (1983). The automatic grammatical tagging of the lob corpus. *ICAME Journal : International Computer Archive of Modern and Medieval English Journal*, **7**, 13–33.

LIGHT M. (1996). Morphological cues for lexical semantics. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, p. 25–31 : Association for Computational Linguistics.

LOFTSSON H. (2009). Correcting a pos-tagged corpus using three complementary methods. In *In Proceedings of EACL-09, 523–531*.

MILNE R. (1986). Resolving lexical ambiguity in a deterministic parser. *Computational Linguistics*, **12**(1), 1–12.

NGAI G. & FLORIAN R. (2001). 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*, p. 1–8 : Association for Computational Linguistics.

ONYENWE I., HEPPLE M., UCHECHUKWU C. & EZEANI I. (2015). 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*, p.̃24.

ONYENWE I. E., UCHECHUKWU C. & HEPPLE M. (2014). Part-of-speech tagset and corpus development for igbo, an african. *LAW VIII*, p.̃93.

PAVEL K. & KAREL O. (2002). (semi-) automatic detection of errors in pos-tagged corpora. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, p. 1–7 : Association for Computational Linguistics.

TALJARD E., FAASS G., HEID U. & PRINSLOO D. J. (2008). On the development of a tagset for northern sotho with special reference to the issue of standardisation. In *Literator 29(1)*.

THEDE S. M. & HARPER M. (1997). Analysis of unknown lexical items using morphological and syntactic information with the timit corpus. In *In Proceedings of the Fifth Workshop on Very Large Corpora*, p. 261–272.

TOUTANOVA K., KLEIN D., MANNING C. D. & SINGER Y. (2003). 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-Volume 1*, p. 173–180 : Association for Computational Linguistics.