

to their agglutinating nature. The vast majority of work has been done in languages with varying systems of inflection, but rarely anything beyond that, often segmenting text before the task of tagging. One shortcoming of software focused on inflecting or ablauting morphology (Yarowsky and Wicentowski, 2000) is the reliance on minimum description length, which is effective for cases of a single affix. Goldsmith (2001) asserts that using algorithms focusing on minimum description length are insufficient, and therefore not useful, in the case of agglutinating languages like Swahili. Hurskainen (1996) cites five major difficulties for morphological disambiguation in Bantu. Four of these relate to the verb. Dealing with verbal morphology, as we do here, is thus an important starting point for empirical morphological analysis for Swahili.

Early experimentation with the Morfessor automated morphological segmentation algorithm (Creutz and Lagus, 2005), intended for languages like Finnish, was not particularly successful. In theory, it should surpass the success of single-affix-focused algorithms. An equal number of segmented morphemes to morphological positions is necessary for the task of tagging or gloss assigning. Using only full verbal data, the algorithm undersegmented; given separate morphological information, the algorithm was over-aggressive in segmentation. Because of this, an alternative to segmentation followed by tagging is necessary for the task at hand.

3 Approach

Classification features in tagging are the syllables of the word itself. The intention is that statistically, the presence of certain syllables would lead to correct assignment of a tag. This leverages the monosyllabicity of Swahili verbal morphemes. By numbering the syllables, a strict order is maintained. This produces three different sets for each experiment: one containing rightward-enumerated syllables, one containing reverse-enumerated syllables, and one with both enumerations.

We focus on the five closed-class morphemes treated in the Helsinki Corpus of Swahili (HCS, 2004) tag set.² The morphemes we tag with glosses

²The full tagset is available in PDF form at <http://www.aakkl.helsinki.fi/comeel/corpus/swatags.pdf>

are as follows:

1. **The Subject Marker (SM).** The SM takes the form of the first leftmost morpheme, and relates to the grammatical subject of the verb. It is in grammatical complementary distribution with infinitive or null markers. There are 32 tags for this position. While many of these tags directly refer to the multiple genders in Swahili, some are also included as being marked for negation, being generally marked for the traditional three-person, singular or plural convention seen in most languages, or simply as being a subject prefix.
2. **The Tense Marker (TM).** The tense-marking morphological position in Swahili accounts for both tense and aspect, and occupies a morphological position immediately following the SM. There are 40 tags in all to consider, many of them being the same tense, but marked for morphological variation, such as the PAST:lisha tag versus the PAST:liisha tag, corresponding to, for example, the difference between *nilishasoma* versus *niliishasoma*, both glossable as ‘I slept’.
3. **The Object Marker (OM).** This marker, directly following the TM, is an optional agreement marker with the grammatical object of the verb. There are 21 tags for this morphological position. 18 correspond directly to gender markers. Two refer to reflexivity. The OBJ tag is a catch-all for any object concord.
4. **The Relative Marker (REL).** Relativization is shown as a marker that refers to the relativized noun phrase. There are 17 separate tags for this position, 15 of which directly refer to a specific gender.
5. **Verbal Extensions (EXT).** Extensions alter the argument structure of a verbal root. It can alter transitivity in several ways as well as the active or passiveness of a verb. Furthermore, it can mark a verb as being stative³ or causative. There are eight tags in the tagset we consider,

³Such as ‘The car is washed’ being in reference to the clean state of the car, rather than the telic point of the action of washing the car.

four of which are morphological variations of the causative marker. For present purposes, we expand the set of extensions into a set of all possible combinations of extensions, as these can be used in conjunction with one another productively to alter semantic meaning as it relates to the verbal root predicate.

An example of the formatted training data for *nilipikiwa* would first be syllabified, and then have the features numerated as per its data set. An example of the the data sets used can be see in Table 3. These features are used to build probabilistic models of $P(\text{gloss}|\text{word})$ using Maximum Entropy and Naive Bayes models.

All three sets capture important optimality-theoretic constraints in their own right. The rightward-enumerated set reflects the constraint ALIGN-X-LEFT, where X refers to each morpheme to be considered.⁴ This constraint reflects the typological proclivity of morphemes to prefix to a root. If we take each syllable to be a possible morpheme, the number for each syllable we see above would be, in fact, $n-1$ violations of this alignment constraint. We therefore encode the syllables in relation to the edge in two manners: the rigid linear ordering found on the surface, and the number of hypothetical violations of the related OT constraint.

The second set mirrors this situation with the constraint ALIGN-X-RIGHT, with the same stipulation as above. This constraint reflects the proclivity of certain morphemes to attach to the rightward periphery of a root. Again, enumeration reflects both rigid backward order with regards to linearity as well as $n-1$ of this particular constraint.

The introduction of the third set, which contains both forward and backward enumerations, functions to normalize both suffix and prefix morphemes; the second syllable from the left may regularly be the tense marker, but this may vary enumerating from the right—and vice versa for verbal extensions. This combination also reflects a crucial component to OT constraints—ranking and interaction. The model has available order and, subsequently, the number of violations for both alignment constraints. Given these,

⁴This is adopted from Kager (1999), which is further an adoption of approaches to Tagalog’s *-um-* morpheme by Prince & Smolensky.

it should be able to outperform a model with either of the previous data sets. Furthermore, its utility should quantitatively affirm the value of optimality-theoretic typological alignment observations, and how they interact in morphological production. This approach is very much in the spirit of work proposed in Johnson and Goldwater (2003) that would use a maximum entropy model rather than OT-specific ranking algorithms.

4 Results

We produce the analysis for an entire word by combining the most probable labels from independent classifiers for each morphological position. Performance was measured by training models on 204,439 word forms and testing on 25,554 *unseen* words. The best performance was a whole-word accuracy of 73.5%, using both feature sets and the maximum entropy tagger. This easily surpassed a baseline of 0.4%⁵ by selecting the value with the highest unigram probability for each position. It also significantly surpassed the Naive Bayes tagger’s best whole-word accuracy. Almost all individual tag accuracies well surpassed their own baselines. The maximum entropy model also generally outperformed its Naive Bayes counterparts at the morpheme level. This can be generally attributed to the non-independence of the features used. Naive Bayes assumes independence of features, while maximum entropy deals well with non-independent feature sets.

Using both feature sets in combination significantly increased the accuracy of either feature set by itself. This quantitatively captures the underlying edge directionality of specific morphemes. Reverse enumeration by itself showed a significant increase in accuracy against forward enumeration for morphemes to the right of the verb root. This also shows this method’s ability to leverage the inherent ordering of extensions as well as their edginess. Because of the synergistic raise in accuracy from the bidirectional feature set, we can assume that edginess constraint interaction can be learned probabilistically.

⁵Due to testing on unique unseen word forms, the baseline for whole words was extremely low. This is a baseline from the most common morpheme for each position. The baseline of the most common word is .7%.

Name	Pattern/Size	Edge	Example (based on <i>nilipikiwa</i>)
L-CV	CV pairs	left	<i>ni.L1 li.L2 pi.L3 ki.L4 wa.L5</i>
R-CV	CV pairs	right	<i>ni.R5 li.R4 pi.R3 ki.R2 wa.R1</i>

Table 1: The two basic feature sets and their composition. An additional set, BI, is the combination of these.

Model	SM	TM	OM	REL	EXT
Bi	88.9	97.1	92.4	95.8	93.8
Rev	85.0	92.1	87.4	95.4	93.0
Fwd	88.4	92.3	88.7	95.2	84.8
BL	15.9	20.0	64.3	76.5	50.2

Figure 1: Maximum entropy morpheme accuracy per feature set compared to the baseline

Set	SM	TM	OM	REL	EXT
Bi	84.9	93.8	81.7	94.4	87.2
Rev	76.7	85.8	77.8	92.8	87.8
Fwd	85.5	92.9	81.0	94.0	79.7
BL	15.9	20.0	64.3	76.5	50.2

Figure 2: Naive Bayes morpheme accuracy per feature set compared to the baseline

Maximum Entropy morpheme-by-morpheme accuracy can be seen in Figure 4. Naive Bayes accuracy is seen in Figure 4. Full-word accuracy is shown in Figure 3. While the best accuracy is still incorrect a quarter of a time, it is rarely totally incorrect. In Figure ??, we show that 4/5ths accuracy, or being incorrect by a single morpheme as opposed to multiple morphemes, is extremely high for the best model. With less than 1/20th of word forms being incorrect by more than one morpheme, this method is still practically viable despite being totally correct for less than 75% of word forms. Because this is testing on totally unseen data, it can be expected that accuracy in a practical application would increase dramatically, following with Zipfian frequency distributions of words.

Model	Bi	Fwd	Rev	Baseline
Maxent	73.5	62.4	63.4	0.4
NB	65.4	58.1	84.4	0.4

Figure 3: Whole word accuracy per feature set and model

Model	Bi	Fwd	Rev
Maxent	95.2	91.4	91.2
NB	85.0	82.9	78.8

Figure 4: Percentage of tagger outputs with a 4/5ths or better accuracy

5 Future Work

Given the success of this tagger, we intend to explore its applicability to the task of speeding up language documentation projects interested in producing interlinearized glossed text. Using syllables as features can serve as a useful method to successfully arrive at proper morphological tags for a variety of languages that follow the typological trait of morpheme-to-syllable correlation. This may make analysis of such languages—especially those that do agglutinate—significantly easier on a linguist in the field. We are investigating the learning curves for these models to determine their suitability for use in semi-automated annotation and active learning to produce interlinear glossed text.

6 Conclusion

The high measure of success above the baseline for all three models show that each data set provides a valuable insight for analysis of verbal morphemes. Using the forward- and backward-enumerated features in a combined feature set serves to make better predictions in the exact areas where either feature set alone would be deficient. This points to the utility of incorporating concepts of both left- and right-edge alignment constraints for the purpose of bootstrapping any statistical morphological analysis algorithm. Furthermore, this can serve as an affirmation that conceptually, edginess is a valuable factor in understanding and learning morphology.

References

E. O. Ashton. *Swahili Grammar*. Longman Group Ltd., Burnt Mill, Harlow, Essex, UK, 1944, 1947, 1984.

- Mathias Creutz and Krista Lagus. Unsupervised morpheme segmentation and morphology induction from text corpora using morfessor 1.0. *Publications in Computer and Information Science*, Report A81, 2005.
- J. Goldsmith. Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27:21–46, 2001.
- HCS. Helsinki corpus of Swahili, 2004. Compilers: Institute for Asian and African Studies (University of Helsinki) and CSC - Scientific Computing.
- A. Hurskainen. Disambiguation of morphological analysis in Bantu languages. In *Proceedings of the 16th Conference on Computational Linguistics*, volume 1, pages 568–573, Copenhagen, Denmark, 1996.
- M. Johnson and S. Goldwater. Learning OT constraint rankings using a maximum entropy model. In *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*, pages 111–120, Stockholm University, Sweden, 2003.
- R. Kager. *Optimality Theory*. Cambridge University Press, 1999.
- A. Prince and P. Smolensky. *Optimality Theory: Constraint Interaction in Generative Grammar*. Blackwell Publishers, 2004.
- D. Yarowsky and R. Wicentowski. Minimally supervised morphological analysis by multimodal alignment. In *Proceedings of ACL-2000*, pages 117–125, Hong Kong, 2000.