Bootstrapping Morphological Analysis of Gĩkũyũ Using Unsupervised Maximum Entropy Learning

Guy De Pauw[1], *Peter Waiganjo Wagacha*[2]

[1]CNTS - Language Technology Group, University of Antwerp, Belgium
[2]School of Computing and Informatics, University of Nairobi, Kenya
guy.depauw@ua.ac.be, waiganjo@uonbi.ac.ke

Abstract

This paper describes a proof-of-the-principle experiment in which maximum entropy learning is used for the automatic induction of shallow morphological features for the resource-scarce Bantu language of Gĩkũyũ. This novel approach circumvents the limitations of typical unsupervised morphological induction methods that employ minimum-edit distance metrics to establish morphological similarity between words. The experimental results show that the unsupervised maximum entropy learning approach compares favorably to those of the established AutoMorphology method.

Index Terms: unsupervised learning, morphology, Bantu languages

1. Introduction

In recent years, research in the field of natural language processing has been marked by a gradual, but definite shift from a deductive, rule-based paradigm to more empirically inspired, inductive approaches. Besides achieving state-of-the-art accuracy on most NLP tasks, these corpus-based methods have the distinct advantage of being inherently portable to other languages, provided there is annotated data available. But while such resources are readily available for commercially interesting languages, the majority of languages in the world can be considered resource-scarce, i.e. lacking digital linguistic resources, as well as lacking the financial resources to create them from scratch.

This not only largely excludes these languages from the prevalent scientific current of corpus-based NLP, but also to some extent serves to widen the digital divide for the native speakers of such under-resourced languages. The lack of effective NLP components establishes an unfortunate paradox in that useful natural language applications, such as machine translation and computer-assisted language learning, cannot be developed for a population that might benefit from them the most.

A recent trend in computational linguistics further amplifies the inductive approach and investigates unsupervised learning techniques that try to induce linguistic classification properties of a language on the basis of unprocessed (i.e. not annotated) text. These language-independent algorithms can provide general linguistic annotation tools with a minimum of manual effort and can therefore bootstrap the development of annotated datasets and NLP tools for resource-scarce languages. On-going research investigates the application of such robust, language-independent machine learning techniques to Kiswahili and local Kenyan languages, like Gĩkũyũ, Dholuo and Kĩkamba [1, 2, 3, 4, 5].

Despite the fact that these techniques hold promising prospects for many smaller sub-Saharan languages, for which vernacular publications are now increasingly being created digitally, most of the current state-of-the-art unsupervised learning algorithms rely on large amounts of digital text data, by definition not available for resource-scarce languages.

This paper describes a proof-of-the-principle experiment with a novel unsupervised learning technique, based on maximum entropy learning, that is able to induce shallow morphological features on the basis of a small lexicon of the resource-scarce Bantu language of Gĩkũyũ. It also improves on current unsupervised morphological induction techniques, by circumventing minimum-edit distance type processing, typically unsuitable for highly inflectional languages like Gĩkũyũ. The machine learning approach described in this paper, is able to establish morphological similarity between words in a more robust fashion, thereby enhancing the induction of its morphological features.

We start off with some background on the target language, Gĩkũyũ, and outline the available corpus material in Section 2. Next, we introduce the novel application of maximum entropy learning as an unsupervised learning technique in Section 3. Section 4 describes the experimental setup, followed by some evaluation results. We finish off with some pointers to future work and concluding remarks.

2. Gĩkũyũ Language

The Gĩkũyũ language is spoken by approximately 6 million Gĩkũyũ people (also referred to as Kĩkũyũ), who are predominantly located around central Kenya. It is a tonal Bantu language and is classified as E31 [6]. It has six dialects, namely Southern Gĩkũyũ (Kiambu, Southern Murang'a), Ndia (Southern Kirinyaga), Gichugu (Northern Kirinyaga), Mathira (Karatina), and Northern Gĩkũyũ (Northern Murang'a, Nyeri). There are several languages that are closely related to it. These include Embu, Chuka, Kĩkamba and Merũ. Like many other languages in Africa, it is digitally resource-scarce due to a competing language environment where official languages English and Kiswahili are the main language of tuition in schools and administration. Today Gĩkũyũ is more spoke than read.

The orthography of Gĩkũyũ further contributes to its digital resource-scarceness, as it incorporates two diacritically marked characters ĩ and ũ representing different phonemes from the unmarked equivalents. OCR methods typically do not perform well on this type of diacritically marked text, rendering automatic corpus collection more troublesome. Furthermore, most of the digital text available, does not include these diacritics. Without an extensive digital word lexicon, alternative methods need to be developed to post-process OCR-ed or web data [3, 5].

The morphology of Gĩkũyũ includes extensive use of modifying (pre)prefixes and verbal subject and object affixation [7]. Like many others, this Bantu language has 17 noun classes. To know what noun class a noun belongs to, is quasi-semantically determined for most classes, as is the case for the human classes 1 (marked by prefix "mũ-") and 2 ("a-"), the infinitive/gerund class 15 ("kũ-") and the locative classes 16 ("ha-") and 17 ("kũ-"). In most cases, the noun class of a word and consequently its semantic properties can to some extent also be deduced by looking at its morphological features, in particular its prefix.

A good morphological analyzer for this language would therefore provide a wealth of syntactic and semantic information. Since lack of commercial and (computational) linguistic interest means that manual effort needs to be kept to a minimum, we attempt an alternative, automatic approach to the induction of morphological features for Gĩkũyũ.

To enable corpus-based research on the Gĩkũyũ language, a 17,000 word corpus was constructed from various types of texts, although with a strong bias towards religious material. The corpus has been compiled by scanning books and web crawling and has been manually corrected and post-processed [2]. The corpus has been uniformly recoded into UTF-8 and tokenized.

The experiments described in this paper were conducted on a 5,000 word lexicon, extracted from this corpus. Note that the size of the corpus and the induced lexicon is well below the recommendations that both LDC and ELRA put forward for dealing with resource-scarce languages. The exploratory experiments described in this paper are geared towards enhancing the current corpus material as a guideline for the development of a more elaborate data set. Furthermore, in the next section we will introduce a novel unsupervised learning technique that is able to overcome some of the limitations of working with such a minimalist data set.

3. (Unsupervised) Maximum Entropy Learning

The unsupervised learning technique proposed in this paper uses the machine learning technique of maximum entropy modeling (maxent) to determine morphological similarity between words. Maxent can be described as a sophisticated statistical method for estimating probability distributions from data. The idea behind maximum entropy is that for some given data, we should prefer the most uniform models that also satisfy any given constraints. When applying the maximum entropy method, the training data is used to set constraints on the conditional probability. Each constraint in turn expresses a characteristic of the training data that should also be present in the induced distribution.

The initial step when using maximum entropy, is to identify a set of feature functions that will be useful for classification. For each of these features, a measure of its expected value over the training data is taken and considered to be a constraint for the model distribution. This guarantees that a unique distribution exists that has maximum entropy. To obtain the parameters of a maximum entropy classifier given a set of constraints, the iterative scaling technique is used to optimize the induced model over a pre-determined number of iterations.

Maximum entropy modeling has consistently achieved top performance on a variety of NLP tasks, such as part-of-speech tagging, prepositional-phrase attachment and text segmentation [1, 8]. Whereas it is most often used as a supervised learning technique, requiring annotated data to build the model, we employ the method to conduct unsupervised learning. The principle however remains the same: we extract features from raw text corpora and consequently build a maximum entropy model. But whereas in supervised learning, one is principally concerned in optimizing the accuracy of a classifier, we are primarily interested in the generalization properties of the induced model.

class	features
ngĩthiĩ	B=n B=ng B=ngĩ B=ngĩt B=ngĩth B=ngĩthi I=g I=gĩ I=gĩt I=gĩth I=gĩthi E=gĩthiĩ I=ĩ I=ĩt I=ĩth I=ĩthi E=ĩthiĩ I=t I=th I=thi E=thiĩ I=h I=hi E=hiĩ I=i E=iĩ
tũgĩthiĩ	B=t B=tũ B=tũg B=tũgĩ B=tũgĩt B=tũgĩth B=tũgĩthi I=ũ I=ũg I=ũgĩ I=ũgĩt I=ũgĩth I=ũgĩthi E=ũgĩthiĩ I=g I=gĩ I=gĩt I=gĩth I=gĩthi E=gĩthiĩ I=ĩ I=ĩt I=ĩth I=ĩthi E=ĩthiĩ I=t I=th I=thi E=thiĩ I=h I=hi E=hiĩ I=i E=iĩ

Figure 1: Example feature values for the words "ngĩthiĩ" (I went) and "tũgĩthiĩ" (we went)

3.1. Morphological Maxent Model

The first step in the induction of morphological features is to determine morphological similarity between words. Like most, if not all, Bantu languages, Gĩkũyũ is highly inflectional, with a large array of affixation possibilities to often short stems. While a rich morphology typically reduces lexical ambiguity [1], it makes the unsupervised induction of morphology problematic: even huge corpora cannot possibly capture all paradigmatic variants of all the word stems. Furthermore, most approaches to unsupervised morphology induction are typically based on some form of minimum-edit distance measuring and are less likely to capture the dependency between morphologically related word like "thoma" (read) and "mũthomere" (manner of reading), since they are orthographically quite different.

We therefore attempt a different approach, in which we measure morphological similarity between words using a morphological maximum entropy model. To this end, we extracted a 5,000 word lexicon from the corpus, containing the unique tokens in the corpus. Given the modest size of the corpus, we did not place any restrictions on the minimal frequency of a token to be included in the lexicon.

Figure 1 illustrates how training instances are extracted. For each word in the lexicon, we list all of its possible orthographic subsets. These subsets constitute the features, while the word itself is considered to be the class to be predicted. There are three types of features: "B="-features describe a subset at the start of the word form, "E="-features indicate patterns at the end of the word and "I="-features describe patterns inside the word form. Note that we are indeed stretching the notion of morphology. A more appropriate term would perhaps be suprasegmental orthography, since at this point, we are merely matching sequences of graphemes, i.e. candidate morphemes.

Many of the features listed in the examples of Figure 1 seem meaningless. "I=h" for example is not likely to contain morphologically relevant information. The general idea however is that a feature like "I=h" will be too common in the training data to provide a useful constraint, whereas a more specialized feature like "I=ngĩthi" (describing the stem as a bound morpheme) might indeed trigger useful morphological generalization properties.

word	predictions
tũgĩthiĩ	ngĩthiĩ (1.78e-07) igĩthiĩ (1.54e-07) agĩthiĩ (1.53e-07) thiĩ (3.52e-08) to (3.28e-09) ti (3.28e-09) ta (3.28e-09) tu (3.28e-09) ĩthiĩ (1.03e-09) tũthiĩ (8.75e-10) magĩthiĩ (8.18e-10) athiĩ (6.93e-10) angĩthiĩ (4.42e-10) tũĩ (2.14e-10) tũu (2.03e-10) ĩgagĩthiĩ (4.48e-11) niĩ (9.81e-12) gũthiĩ (7.78e-12) ...

Figure 2: Morphological Maxent Model output for the word "tũgĩthiĩ"

The lexicon is thus transformed into a training set of 5,000 instances. This data is then used to train a maximum entropy classifier [9] until maximum classification accuracy on the training set is reached. Rather than using the resulting model to classify new data, we use it to re-classify the training data itself. For each word in the training data, we provide the *n* most likely classes according to the induced model. In this case this outputs the morpho-orthographically most similar words and a probability value expressing the degree of similarity (Figure 2).

While the output does not yet provide true morphological clustering at this point, it is able to find relationships between morphologically related words, that a typical minimum edit distance approach would ignore (for instance between "tũgĩthiĩ" and "igagĩthiĩ" or between "thoma" and "mũthomere"). Previous experiments [4] showed that this output, combined with a simple clustering technique, can provide relevant morphological subclasses of words.

Note that we are employing maximum entropy learning as a way of establishing morphological similarity between the words of the lexicon and that we are not trying to optimize any kind of classification accuracy. Words are classified as being morphologically similar to other words within the same data set. Reclassifying the training data itself therefore does not provide any type of unfair advantage. While one would indeed typically expect a strict division between training set and test set in a machine learning experiment, this point is moot for the type of processing described in this paper.

3.2. Prefix Extraction

Using the groups of morphologically related words, output by the maxent model, we can now start identifying how exactly these words are related, be it through a similar stem, suffix and/or prefix. In this paper, we describe how we can extract a list of possible (pre)prefixes for the Gĩkũyũ language.

The algorithm to identify the possible prefixes directly process the output of the maxent model (Figure 2). For each line of the output, the token in the *word* field is extracted. For this word, we then consider all possible combinations of prefix and "stem"[1]. This is exemplified in Figure 3 for the word *tũgĩthiĩ*.

For each candidate prefix, we then try to pattern match the resulting stem with each of the tokens predicted by the maxent model (column *predictions* in Figure 2). If the stem matches the word, the score of the candidate prefix is equal to the associated -log probability output by the maxent model. For each candidate prefix, the cumulative score is maintained in a table.

The candidate prefix "tũ-" in Figure 3 for example (which in this case is the correct one) yields the stem gĩthiĩ. This stem matches a fair amount of tokens considered by the maxent classifier to be morphologically related, three of which are listed in the column *matches* in Figure 3. After pattern-matching with these three tokens, its cumulative score is 23.8.

[1]In this experiment, we consider the stem as the combination of stem, suffixes and possible infixes.

prefix	stem	matches	score
0	tũgĩthiĩ	—	–
t	ũgĩthiĩ	—	–
tũ	gĩthiĩ	ngĩthiĩ	15.5
		igĩthiĩ	15.7
		ĩgagĩthiĩ	23.8
		...	...
tũg	ĩthiĩ	ĩthiĩ	20.7
		magĩthiĩ	20.9
		angĩthiĩ	21.5
		...	...
tũgĩ	thiĩ	ngĩthiĩ	15.5
		thiĩ	17.1
		athiĩ	21.1
		...	...
tũgĩt	hiĩ	thiĩ	17.1
		tũthiĩ	20.9
		gũthiĩ	25.6
		...	...
tũgĩth	iĩ	ĩthiĩ	20.7
		angĩthiĩ	21.5
		niĩ	25.3
		...	...
tũgĩthi	ĩ	ngĩthiĩ	15.5
		niĩ	25.3
		gũthiĩ	25.6
		...	...

Figure 3: (Summarized) sample output of prefix candidate scoring.

After all the lines of the maxent output file are processed, the prefixes, ranked according to their score, are output. Experimental results (Section 4) show that this approach, although rudimentary, is quite effective at identifying the list of possible prefixes of the Gĩkũyũ language.

4. Experimental Results

In this section, we describe a proof-of-the-principle experiment in which we extract a list of prefixes for the Gĩkũyũ language. We compare the output of our approach with that of the established AutoMorphology package described in [10].

AutoMorphology is a software package that tries to automatically induce the morphological features of a natural language. Excellent results have been reported for the automatic induction of the morphology of Indo-European languages, with the output of the program matching that of a human morphologist. By the author's own admission, the package is less suitable for non-Indo-European languages, as it, among other issues, presupposes that there are no more than two prefixes attached to any given stem. As the only available package for

Algorithm	Precision	Recall	$F_{\beta=1}$
AutoMorphology	68.4%	37.1%	48.1%
Unsupervised Maxent	70%	60%	64.6%

Table 1: Experimental Results for Prefix Retrieval Experiment

unsupervised morphology induction, it nevertheless provides a good baseline against which we can compare the unsupervised maxent approach.

The approach presented in this paper, departs from an even more rigid assumption than the AutoMorphology package, as it attempts to extract one single prefix group, possibly consisting of several adjoined (pre)prefixes. Experimental results however show that our approach is able to induce prefixes that AutoMorphology fails to detect, indicating the prefix restriction is not the only issue biasing it towards Indo-European types of morphology.

We used the same dataset to train both the unsupervised maximum entropy learning approach described in this paper and the AutoMorphology method. Since the latter is not UTF-8 compatible, some minor data conversion needed to be done: ĩ and ũ were consistently replaced by Latin characters x and q, which are not present in Gĩkũyũ orthography.

Since there is not yet a gold-standard morphologically annotated corpus for Gĩkũyũ, we opted for a more qualitative evaluation, by scoring the algorithm in terms of precision (how many of the predicted prefixes are correct) and recall (how many of the prefixes in Gĩkũyũ are retrieved).

Table 1 shows the results of this experiment. AutoMorphology is careful in its prediction of morphemes: it predicts 19 morphemes, 13 of which are correct (equaling to a precision of 68.4%). With 35 prefixes to be predicted, this yields a recall score of just 37.1%. The unsupervised maximum entropy learning approach predicts 30 prefixes, 21 of which are correct, thereby improving on both Automorphology's precision and recall scores. These results are encouraging as they indicate that the unsupervised maximum entropy learning approach is able to retrieve a fairly accurate list of prefixes for a very limited data set of a highly inflectional and resource-scarce Bantu language.

5. Future Work and Conclusion

In this paper we introduced a novel unsupervised learning technique using a maximum entropy backbone to describe morphological similarity between words. The proposed method has the distinct advantage that it is robust to modest size corpora. Furthermore, the maximum entropy classification was able to capture morpho-orthographical dependencies that a typical minimum-edit distance would ignore. We described how the probabilities output by the maximum entropy model can aid the automatic extraction of prefixes of a resource-scarce language.

While the functionality of the unsupervised maximum entropy learning approach is not as extensive as that of the established AutoMorphology method, the latter's bias to Indo-European languages seems to put it at a disadvantage when applied to the Bantu language of Gĩkũyũ. The experiments described in this paper yield some preliminary, but nevertheless quite encouraging results.

Future research efforts in this area will concentrate on the extraction of all types of preprefixes, suffixes and infixes, as well as the automatic retrieval of stems. The proposed method will also be applied to other African languages such as Dholuo, Kĩkamba, Kinyarwanda and Kiswahili. Not a resource-scarce language, the latter will give us an idea of the scalability of the approach. Furthermore, it can be evaluated using the gold-standard lemmatization available in the Helsinki Corpus of Swahili [11]. This will allow us to quantitatively evaluate our approach and its apparent attractive properties in the processing of highly inflectional languages and Bantu languages in particular.

6. Acknowledgments and Demo

The research presented in this paper was made possible through the support of the VLIR-IUC-UON program. The first author is funded as a Postdoctoral Fellow of the Research Foundation - Flanders (FWO). A demonstration system of the unsupervised morphology induction approach described in this paper is available on http://aflat.org.

7. References

[1] G. De Pauw, G.-M. de Schryver, and P. Wagacha, "Data-driven part-of-speech tagging of Kiswahili," in *Proceedings of Text, Speech and Dialogue, 9th International Conference*, ser. Lecture Notes in Computer Science, P. Sojka, I. Kopeček, and K. Pala, Eds., vol. 4188/2006. Berlin, Germany: Springer Verlag, 2006, pp. 197–204.

[2] P. Wagacha, G. De Pauw, and K. Getao, "Development of a corpus for Gĩkũyũ using machine learning techniques," in *Proceedings of LREC workshop - Networking the development of language resources for African languages*. Genoa, Italy: ELRA, 2006, pp. 27–30.

[3] P. Wagacha, G. De Pauw, and P. Githinji, "A grapheme-based approach for accent restoration in Gĩkũyũ," in *Proceedings of the Fifth International Conference on Language Resources and Evaluation*. Genoa, Italy: ELRA, 2006, pp. 1937–1940.

[4] G. De Pauw, P. Wagacha, and D. Abade, "Unsupervised induction of Dholuo word classes using maximum entropy learning," in *Proceedings of the First International Conference in Computer Science and Informatics (COSCIT 2007)*. Nairobi, Kenya: University of Nairobi, 2007.

[5] G. De Pauw, P. Wagacha, and G.-M. de Schryver, "Automatic diacritic restoration for resource-scarce languages," in *Text, Speech and Dialogue. Proceedings of the 10th International Conference TSD 2007*, ser. Lecture Notes in Computer Science, V. Matousek and P. Mautner, Eds. Berlin, Germany: Springer Verlag, 2007 (forthcoming).

[6] M. Guthrie, *The Classification of Bantu Languages*. London: Dawsons of Pall Mall, 1967.

[7] J. M. Mugane, *A Paradigmatic Grammar of Gĩkũyũ*. Stanford California: CLSI Publications, 1997.

[8] A. Ratnaparkhi, "Maximum entropy models for natural language ambiguity resolution," Ph.D. dissertation, University of Pennsylvania, 1998.

[9] Z. Le, "Maximum entropy modeling toolkit for python and c++," Centre for Speech Technology Research of the University of Edinburgh, Tech. Rep., 2004.

[10] J. Goldsmith, "Unsupervised learning of the morphology of a natural language," *Compututational Linguistics*, vol. 27, no. 2, pp. 153–198, 2001.

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