Unsupervised Induction of Dholuo Word Classes using Maximum Entropy Learning

Guy De Pauw[1], Peter W. Wagacha[2] and Dorothy Atieno Abade[2]

[1] CNTS - Language Technology Group, University of Antwerp (Belgium)
guy.depauw@ua.ac.be
[2] School of Computing and Informatics, University of Nairobi (Kenya)
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 word classes for the Western Nilotic language of Dholuo. The proposed approach extracts shallow morphological and contextual features for each word of a 300k text corpus of Dholuo. These features provide a layer of linguistic abstraction that enables the extraction of general word classes. We provide a preliminary evaluation of the proposed method in terms of language model perplexity and through a simple case study of the paradigm of the verb stem "somo".

1 Introduction

Linguistic description typically involves finding some formal layer of abstraction that is able to capture the general (phonological, morphological, syntactic, ...) patterns underlying a language. While this is traditionally done by hand in theoretical linguistics, corpus-based computational approaches can provide an interesting complementary alternative that is purely empirical in nature and therefore less susceptible to the pitfalls of theory dependent interpretation. While these data-driven approaches have been well-studied for Indo-European languages, their reliance on large volumes of (annotated) data make them a less obvious choice for the linguistic description of African languages, many of which are resource-scarce.

This paper describes a novel computational approach to corpus-based linguistic classification: an inherently language-independent clustering technique based on maximum entropy learning that is able to induce general word classes from a relatively modest collection of raw text data for the resource-scarce language of Dholuo.

2 Linguistic classification from scratch

In recent years, research in the field of computational linguistics and natural language processing (NLP) has been marked by a gradual, but definite shift from a deductive, rule-based paradigm to more empirically inspired, inductive approaches. Rather than hard coding the solution

to a particular NLP problem in a set of hand-crafted rules, the latter data-driven methods try to extract the required linguistic classification properties from large, annotated corpora of natural language.
A recent trend in computational linguistics further amplifies the inductive approach and investigates unsupervised learning algorithms for the automatic induction of linguistic classification. Unsupervised learning techniques have the distinct advantage that they can be applied to unprocessed (i.e. not annotated) language data to provide a general linguistic annotation, which can bootstrap corpus-based linguistic research and the development of NLP tools for these 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].
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 clustering technique, based on maximum entropy learning, that is able to induce word classes on the basis of a modest-size text corpus of Dholuo.

3 Dholuo

Dholuo is spoken by the Luo people on the eastern shores of Lake Victoria in both Kenya and Tanzania with about 3.5 million native speakers and is classified as a western Nilotic language. Kenya has two official languages, English and Kiswahili, and over forty different local languages associated with diverse ethnic groups. Whereas in South Africa there is a language policy that enshrines these local languages in the constitution, this is not the case in Kenya. A few decades ago, basic primary education was still being taught in mother tongue in rural areas. For political reasons Kiswahili is now projected as the default language of communication, while English is being used in educational contexts. This has contributed to Dholuo, like all other local languages, being resource scarce.
This restrictive language use practice is also responsible for a waning competence among the younger generations in their own mother tongue. This is to some extent countered by an increased country-wide interest in local languages, as can be clearly noted from the high number of vernacular radio broadcasts and increasing periodic publications.
Dholuo is closely related to Acholi and Lango, spoken in Uganda. Dholuo has two dialects, namely the Trans-Yala dialect spoken in Ugenya, Alego, Yimbo and parts of Gem; and the South Nyanza dialect spoken in South Nyanza, Siaya and Kisumu. Although these dialects are mutually intelligible, they are distinct enough to enable one to tell where the particular speaker comes from, or at least to which group one officially belongs to, merely by his/her accent [4]. Officially there exists no standard dialect.
Like Kiswahili, the Dholuo language uses the Latin alphabet without extra diacritics. The tonality of Dholuo is not marked in the orthography.

There are 26 consonants and nine vowels. The vowels are distinguished primarily by [ATR] harmony [4,5], but the Dholuo orthography does not adequately differentiate these vowels, establishing significant disambiguation problems for humans and automatic methods alike.
To enable corpus-based research on the Dholuo language, a corpus of Dholuo has been compiled. The corpus used for the experiments described in this paper, consists of about 300,000 words 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 post-processed. The corpus has been uniformly recoded into UTF-8 and has been tokenized.

4 Maximum Entropy Learning

The method proposed in this paper uses the machine learning technique of maximum entropy modeling (maxent) to induce word classes. 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 [6]. 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 more interested in the generalization properties of the induced model.

4.1 Morphological Maxent Model

The first model tries to extract word classes on the basis of their morphology, i.e. internal word structure. To this end, we extracted a 26,000 word lexicon from the 300,000 word corpus of Dholuo. For each word in the lexicon, we consider its possible orthographic subsets. If a subset matches another token in the lexicon, it is used as a possible informative

morphological feature. This is illustrated in Example 1 for the words "andiko" and "gindiko".

Example 1

andiko	I=n I=ndi E=iko B=an I=k I=d E=o I=di E=ndiko B=andik B=a I=ik I=nd I=ndik I=i E=ko
gindiko	I=n I=ndi B=g E=iko B=gin I=k I=d B=gi E=o I=di E=ndiko I=in I=ik I=nd I=ndik I=i I=indik E=ko

The first item is the class to be predicted (in this case the complete word), followed by the morphological features describing this word. There are three types of features: "B="-features describe a matching pattern at the start of the word form, "E="-features describe matching patterns at the end of the word and "I="-features describe matching patterns inside the word form.

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

The lexicon is transformed into a training set of 26,000 instances. This data is then used to train a maximum entropy classifier [7] over thirty iterations. 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 (log) probability value expressing the degree of similarity:

Example 2

gihinyogo gihiny 18 ginyodhe 19 gibuogo 20 diyogo 20 kanyogo 21

While the output does not yet provide true morphological clustering at this point, it is able to find relationships that a typical minimum edit distance approach would ignore (for instance the similarity between the related words of "kanyogo" and "gihinyogo"). In Section 5 we describe how we can use this information to extract morphological word classes.

4.2 Contextual Maxent Model

The second model describes words in their sentence context, as observed in the corpus. For this model, we create an instance for each word in the 300,000 word corpus, again using the word as the class to be "predicted", but this time using the surrounding words as features. This is illustrated in Example 3.

Example 3

wabiro	W-2=to W-1=kaka W+1=bedo W+2=achien

Similar to the morphological model, we re-classify the training data and provide the n most likely words. For the instance in Example 3, the model provides the output in Example 4. As opposed to the morphological model, this output does not describe orthographically similar words, but rather words that are likely to appear in a similar sentence context.

Example 4
wabiro mar nyalo joma ng'ama kuom ne onyalo en koro

In the next section, we conduct some experiments in which we try to exploit the output of this model in a practical context.

5 Experiments

The output of the maxent models can be compiled into large networks of interconnected nodes, one for the morphological model and one for the contextual model. Each node in a network describes a word in the lexicon, while edges between nodes denote a morphological or contextual relationship. This network provides valuable information that can be used by clustering algorithms to extract classes of words on the basis of their morphological or contextual properties. Since we do not yet have a gold-standard against which we can evaluate our method, we have conducted some proof-of-the-principle experiments that provide a preliminary evaluation. We first attempt a quantitative evaluation by measuring the reduction of language model perplexity using the induced word classes, followed by a more qualitative evaluation of a specific case study.

5.1 Language Model Perplexity

Using a very simple clustering technique, we can extract word classes from the networks output by the maximum entropy models. To do this, we first order all of the links in the network, according to the probabilities output by the maximum entropy models. Starting with the strongest link, we then consider for each link the node (word) with the highest frequency in the corpus and extract a subnetwork for that node up to a predefined depth. An example of such a subnetwork, centered around the word "misomo" is displayed in Figure 1. The subnetworks are consequently considered as clusters. Once a word belongs to a particular cluster, it is unavailable to other clusters. While this is an overly greedy approach to clustering, it allows us to easily evaluate the output of the maximum entropy models in a language modeling experiment.

The idea in language modeling is to attach a probability to each word in a sentence. This is typically done using n-gram models, which calculate the probability of a word on the basis of the n preceding words. Language modeling is useful in applications such as speech recognition and machine translation, where we want a measure of the likelihood of a certain word appearing given some linguistic context. By looking at the average probability with which a token is predicted, we get an idea of how well the language model fits the data. This is described in the perplexity value: the lower the perplexity, the better the language model is able to "encode" the data.

When clusters, de facto linguistic categories, are available, we can translate each word in the corpus into its cluster class. This reduces the number of unique tokens. If the clusters relate to linguistically relevant categories, the language model has an easier time encoding the data, thereby reducing perplexity.

Method	Perplexity
CMU trigram Random Clusters	203.86
CMU trigram Morph. Clusters (MED)	180.23
CMU trigram Morph. Clusters (MAX)	172.52
CMU trigram Context Clusters (MAX)	167.43

Table 1. Language Model Perplexity

For these experiments we used the CMU - Language Modeling Toolkit [8]. We divided the corpus into a 90% training set, used to train the language model, and a 10% test set on which we measure perplexity. The experimental results can be found in Table 1. The first baseline for this experiment is a random clustering technique which puts the words in 5000 random clusters. This language model achieves a perplexity of 204 on the test set. The second baseline creates 5000 clusters by grouping words according to their Levenshtein (minimum edit) distance [9]. This language model establishes a significant reduction of perplexity over the random clustering baseline.
A trigram language model using 5000 clusters extracted from the network created by the morphological maxent model, further reduces perplexity to 173. This provides some evidence that the clusters created by our approach outperform those of the typically used minimum edit distance method. The contextual clustering technique, which is specifically geared towards capturing local contextual information, further reduces perplexity to 167. While these results by no means establish a definite quantitative evaluation of the linguistic relevance of the induced clusters, they do provide some evidence that the method yields a useful layer of linguistic abstraction.

5.2 Case Study: the verbal paradigm of "somo"

Dholuo has a strongly agglutinating morphology. While this typically reduces lexical ambiguity [1], it makes the unsupervised induction of morphology problematic. Even huge corpora cannot possible capture all paradigmatic variants of all the word stems. Furthermore, typical approaches to unsupervised morphology induction are based on minimum-edit distance and are less likely to capture the dependency between for instance "kanyogo" and "gihinyogo", since they are orthographically quite different. While our clustering approach is also limited by the data-sparsity problem, the decrease in perplexity compared to a minimum edit distance approach (Table 1) indicate it to be more robust towards the latter problem.
Figure 1 displays a subnetwork of the morphological network, centered around the word "misomo", with the length of the edges to some extent encoding morphological similarity. For reasons of space, we limited the depth of the displayed subnetwork. Increasing the depth of the subnetwork allows us to capture all instantiations of the verbal paradigm of the

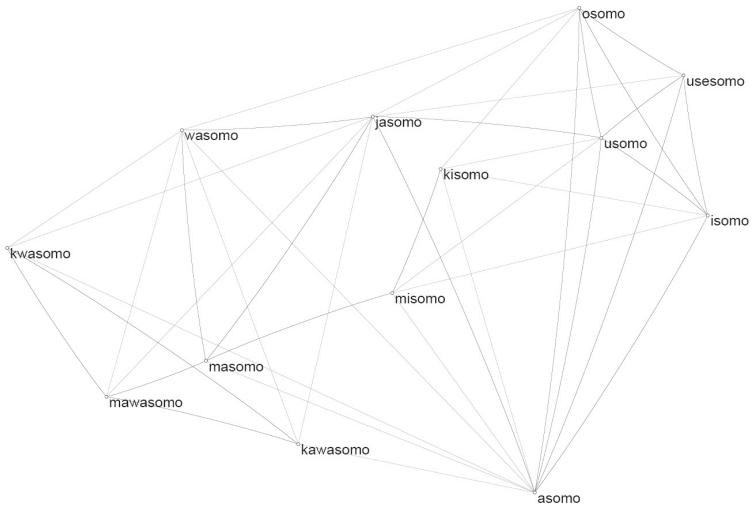

Fig. 1. Morphological Subnetwork around "misomo"

verb stem "somo" present in the corpus [10]. While increasing the size of the subnetwork also increases the number of "false positives", it also allowed us to identify a possible typo in the corpus, namely "wasomoe". Coupled with a network visualization tool, our method can provide an interesting tool for descriptive linguists who want to evaluate their insights against data in a language corpus.

6 Future Work and Conclusion

In this paper we introduced a novel clustering technique using a maximum entropy backbone to describe morphological and contextual relationships 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.
The experiments described in this paper yield some very preliminary, but already quite promising results. Future research efforts in this area will now concentrate on optimizing the information contained in the resulting networks. Whereas the edges are now simply labeled with the similarity value between nodes, we are aiming at providing more descriptive labels that not only describe the strength of the relationship, but also the nature of the similarity. This includes a description of the type of affix or stem that connects the words.
An important limitation to our experiments was the overly greedy clustering technique used to extract word categories from the network. Future experiments will employ more intricate clustering techniques. Combined with more informative edge labels, these should enable the induction of true morphological information.

The proposed method will also be applied to other African languages such as Gĩkũyũ, Kĩkamba, Kinyarwanda and Kiswahili. The latter 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 languages with an agglutinating morphology.

References

1. De Pauw, G., de Schryver, G., Wagacha, P.: Data-driven part-of-speech tagging of Kiswahili. In Sojka, P., Kopeček, I., Pala, K., eds.: Proceedings of Text, Speech and Dialogue, 9th International Conference. Volume 4188/2006 of Lecture Notes in Computer Science., Berlin, Germany, Springer Verlag (2006) 197–204
2. Wagacha, P., De Pauw, G., Getao, K.: Development of a corpus for Gĩkũyũ using machine learning techniques. In Roux, J., ed.: Proceedings of LREC workshop - Networking the development of language resources for African languages, Genoa, Italy, European Language Resources Association, ELRA (May, 2006 2006)
3. Wagacha, P., De Pauw, G., Githinji, P.: A grapheme-based approach for accent restoration in Gĩkũyũ. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation, Genoa, Italy, European Language Resources Association, ELRA (May, 2006 2006) 1937–1940
4. Okoth-Okombo, D.: Dholuo Morphophonemics in a Generative Framework. Dietrich Reimer Verlag, Berlin, Germany (1982)
5. Omondi, L.: The major syntactic structures of Dholuo. In Heine, B., Mohlig, W., Rottland, F., eds.: Language and Dialect Atlas of Kenya (suppl.1). Dietrich Reimer Verlag, Berlin, Germany (1982)
6. Ratnaparkhi, A.: Maximum Entropy Models for Natural Language Ambiguity Resolution. PhD thesis, University of Pennsylvania (1998)
7. Le, Z.: Maximum entropy modeling toolkit for python and c++. Technical report, Centre for Speech Technology Research of the University of Edinburgh (2004)
8. Clarkson, P., Rosenfeld, R.: Statistical language modeling using the CMU–cambridge toolkit. In: Proceedings of Eurospeech '97, Rhodes, Greece (1997) 2707–2710
9. Levenshtein, V.: Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady **10(8)** (1966) 707–710
10. McCall, B.: Luo verb morphology: a description and an optimality-theoretic account (1999) Available from www.pewtergallery.com/betsy//writing/academic/luo/verbs.html.
11. Hurskainen, A.: HCS 2004 – Helsinki Corpus of Swahili. Compilers: Institute for Asian and African Studies (University of Helsinki) and CSC (2004)