Resource-Light Bantu Part-of-Speech Tagging

Guy De Pauw[†], Gilles-Maurice de Schryver[‡], Janneke van de Loo[†]

[†]CLiPS - Computational Linguistics Group
University of Antwerp
Antwerp, Belgium
guy.depauw@ua.ac.be
janneke.vandeloo@ua.ac.be

[‡]Dept of Languages and Cultures
Ghent University
Ghent, Belgium
gillesmaurice.deschryver@UGent.be

[‡]Xhosa Dept
University of the Western Cape
Cape Town, South Africa

Abstract

Recent scientific publications on data-driven part-of-speech tagging of Sub-Saharan African languages have reported encouraging accuracy scores, using off-the-shelf tools and often fairly limited amounts of training data. Unfortunately, no research efforts exist that explore which type of linguistic features contribute to accurate part-of-speech tagging for the languages under investigation. This paper describes feature selection experiments with a memory-based tagger, as well as a resource-light alternative approach. Experimental results show that contextual information is often not strictly necessary to achieve a good accuracy for tagging Bantu languages and that decent results can be achieved using a very straightforward unigram approach, based on orthographic features.

1. Introduction

Part-of-speech tagging is often considered as a prototypical classification task in the field of Natural Language Processing (NLP). It can more generally be described as *sequence tagging* and methods suitable for part-of-speech tagging can be directly applied to a wide variety of other NLP tasks, such as word sense disambiguation (Veenstra et al., 1999), phrase chunking (Ramshaw and Marcus, 1995) and concept tagging (Hahn et al., 2008). It is commonly accepted that contextual features constitute the most important information source to trigger the correct sequence tag of ambiguous words, although for the task of part-of-speech tagging (pseudo-)morphological features are often used as well (Ratnaparkhi, 1996; Daelemans et al., 2010).

In the English sentence in Example 1, part-of-speech disambiguation of the token *can*, which can be a modal, verb or noun, can be performed on the basis of the preceding determiner.

(1) *The can is empty* .

This disambiguation task can be automatically induced on the basis of annotated data: a statistical technique or machine learning algorithm observes the manually annotated data and automatically identifies the most important linguistic features towards disambiguation. This approach has the advantage of being language independent: all that is needed, is annotated data in the target language.

While these *data-driven* approaches have yielded state-of-the-art part-of-speech tagging accuracies for a number of sub-Saharan African languages, such as Swahili (De Pauw et al., 2006), Amharic (Gambäck et al., 2009), Wolof (Dione et al., 2010) and Northern Sotho (de Schryver and De Pauw, 2007), no research efforts exist that explore which type of linguistic features contribute to accurate part-of-speech tagging for the languages under investigation. This paper describes feature selection experiments with a memory-based tagger, as well as a resource-light alternative approach. Experimental results show that contextual information is often not strictly necessary to achieve a good accuracy for tagging Bantu languages and that decent results can be achieved using a very straightforward unigram approach, based on orthographic features.

This paper is organized as follows: in Section 2, we describe some of our previous research efforts on part-of-speech tagging of Bantu languages and we introduce the data sets that were used in the experiments described in this paper. Section 3 outlines experimental results using an off-the-shelf data-driven tagger and discusses the automatically determined optimal combination of features for disambiguation. We introduce a new data-driven approach to part-of-speech tagging of Bantu languages in Section 4 and further contrast it with the traditional, context-driven approach by means of learning curve experiments in Section 5. We conclude with a discussion of the main insights gained from these experiments and pointers for future research in Section 6.

2. Part-of-Speech Tagging for Bantu Languages

In this paper, we will investigate two different approaches to data-driven part-of-speech tagging for four different languages. The languages under investigation are: (i) Swahili, spoken by over fifty million people in eastern Africa, especially in Tanzania and Kenya, (ii) Northern Sotho and (iii) Zulu, two of the eleven official languages in South Africa, spoken by respectively four and ten million people, and (iv) Cilubà, spoken by six million people in the Democratic Republic of the Congo. There is no ideological reason behind the selection of these languages, other than the fact that they are among the few Sub-Saharan African languages that have part-of-speech tagged data available to them.

All four happen to be Bantu languages, which is a subgroup of one of Africa's four language phyla, Niger-Congo. The term sub-group is an understatement, as genetically the classification path goes through Niger-Congo > Mande-Atlantic-Congo > Ijo-Congo > Dogon-Congo > Volta-Congo > East Volta-Congo > Benue-Congo > East Benue-Congo > Bantoid-Cross > Bantoid > South Bantoid > Narrow Bantu. The Narrow Bantu languages, more of-

	Swahili	Northern Sotho	Zulu	Cilubà
Number of sentences	152,877	9,214	3,026	422
Number of tokens	3,293,955	72,206	21,416	5,805
POS-Tag set size	71	64	16	40
% of ambiguous words	22.41	45.27	1.50	6.70
Average % of unknown words	3.20	7.50	28.63	26.93

Table 1: Quantitative information for Swahili, Northern Sotho, Zulu and Cilubà data sets.

ten simply referred to as 'the Bantu languages', are thus truly a late offshoot historically speaking, and still there are about 500 Bantu languages, the largest unitary group on the African continent.

Data Sets

For Swahili, we will use the Helsinki Corpus of Swahili (Hurskainen, 2004a) as our data set of choice. The Helsinki Corpus of Swahili has been automatically annotated using SALAMA, a collection of finite-state NLP tools for Swahili (Hurskainen, 2004b). It is therefore important to point out that this resource constitutes silver-standard data, rather than gold-standard (i.e. manually annotated) data. Previous research (De Pauw et al., 2006) investigated the applicability of four different off-the-shelf, data-driven part-of-speech taggers and various system combination techniques, yielding an overall tagging accuracy of up to 98.6%. In this paper, we will use the same, cleaned-up version of the Helsinki Corpus of Swahili, containing over three million tokens, as described in De Pauw et al. (2006).

For Northern Sotho, we use a gold-standard, manually annotated part-of-speech tagged corpus, first described in de Schryver and De Pauw (2007). This corpus was used as training data for a maximum-entropy based tagger, which is able to deal with pseudo-morphological information that is more suitable to tagging a Bantu language, compared to existing off-the-shelf taggers. de Schryver and De Pauw (2007) describe experiments using this small data set of a little over 10,000 tokens and report a more than encouraging tagging accuracy of 93.5%. Since then, additional data has been automatically annotated by this tagger and manually corrected, yielding a modestly-sized, gold-standard corpus of about 70,000 tokens that will be used during the experiments described in this paper.

A publicly available, but at 21,000 tokens modestly sized, part-of-speech tagged corpus of Zulu is described in Spiegler et al. (2010). This data set will be included in the experiments as well.

Finally, a very small annotated corpus for Cilubà of just 6,000 tokens was manually annotated. While this data set is clearly too diminutive as training material for off-the-shelf data-driven taggers, experiments will show that the Bag-of-Substrings approach (cf. Section 4) is able to perform fairly accurate tagging of unknown words on the basis of a limited amount of training data.

Table 1 shows some quantitative information for the four data sets under investigation, such as the number of sentences and tokens in the first two lines. The third line displays the number of distinct POS tags, used in the corpus. As a rule of thumb, the larger the tag set, the more fine-grained the morpho-syntactic description and consequently the more difficult the task of part-of-speech tagging.

The fourth line in Table 1 expresses the percentage of words in the corpus that are lexically ambiguous, i.e. have been observed with more than one tag[1]. A low lexical ambiguity rate typically indicates a fairly straightforward part-of-speech tagging task. For the Zulu data set, for instance, 98% of the words do not need disambiguation. This is normal for languages which have both a rich morphology and are written conjunctively: here a token typically has considerable affixation, encoding its morpho-syntactic (and often also semantic) properties and is therefore less likely to be lexically ambiguous. This does not hold, of course, for languages with a disjunctive writing system, such as Northern Sotho, as is apparent from its lexical ambiguity rate in Table 1. The degree of conjunctiveness / disjunctiveness (Prinsloo and de Schryver, 2002) for Cilubà lies in-between that of Zulu and Northern Sotho, as does its lexical ambiguity rate.

The last line of Table 1 indicates the average expected number of unknown words in unseen data. For the larger data sets (Swahili and Northern Sotho), this percentage is reasonably low. For the smaller data sets (Zulu and Cilubà), however, this equals to more than one out of four unknown tokens in a typical data set. In Section 4, we will describe a novel approach that is able to classify unknown words with a higher degree of accuracy than a traditional, context-driven tagging method.

3. Context-Driven Tagging

Data-driven taggers have become staple tools in the field of Natural Language Processing and a wide range of different implementations, using a variety of machine learning and statistical techniques, are publicly available. For the world's most commercially interesting languages, these methods are well researched. For Sub-Saharan African languages however, they have only recently been applied. In this section, we will describe experiments with MBT, a data-driven tagger that uses a memory-based learning classifier as its backbone (Daelemans et al., 2010).

Memory-Based Tagging

MBT uses annotated data to automatically induce a tagger that is able to classify previously unseen sentences. To this end, it actually builds two separate memory-based learning classifiers: one for known words (i.e. words that have

[1] As a point of reference: the English Wall Street Journal Corpus (Marcus et al., 1993) contains about 35% lexically ambiguous words.

	Swahili		
	Known	**Unknown**	**Total**
Baseline	97.7	73.37	96.92
	±0.03	±0.37	±0.02
MBT	**98.43**	89.81	98.16
	±0.03	±0.59	±0.04
BoS	97.4	**93.52**	97.27
	±0.03	±0.28	±0.03
Best Combo	98.43	93.52	**98.27**

	Northern Sotho		
	Known	**Unknown**	**Total**
Baseline	90.38	61.81	88.24
	±0.44	±2.53	±0.45
MBT	**95.73**	81.72	94.68
	±0.27	±1.86	±0.24
BoS	85.71	**84.1**	85.59
	±0.38	±1.69	±0.34
Best Combo	95.73	84.1	**94.86**

	Zulu		
	Known	**Unknown**	**Total**
Baseline	99.65	52.02	86.00
	±0.16	±2.38	±1.13
MBT	99.61	75.51	92.71
	±0.19	±1.74	±0.66
BoS	**99.65**	**83.32**	**94.97**
	±0.16	±1.47	±0.50
Best Combo	99.65	83.32	94.97

	Cilubà		
	Known	**Unknown**	**Total**
Baseline	95.72	48.85	82.77
	±1.85	±5.72	±2.24
MBT	**95.91**	62.13	86.58
	±1.66	±5.66	±2.7
BoS	95.84	**72.24**	89.32
	±1.70	±5.56	±2.69
Best Combo	95.91	72.24	**89.37**

Table 2: Results (tagging accuracy & standard deviation (%)) of ten-fold cross validation experiment for Swahili, Northern Sotho, Zulu and Cilubà data sets.

previously been encountered in the training corpus) and another classifier that is able to tag unknown words. In a first phase, MBT constructs for each token in the training corpus a so-called `ambitag`, a single token encoding the tags that have been associated with that word in the training corpus. A word such as "*can*" for example, may receive an ambitag `MD_NN_VB`, which means that it has been encountered as a modal, a noun and a verb.

The training data is then transformed into a series of vectors, describing each word in its context. The known-words classifier uses the `ambitag` of each word as its primary feature. The user can then add extra contextual information: the tags of the left context of the word can be added to the vector to allow for disambiguation of cases such as "*can*" in Example 1. Also the right context of the token can be added, but since tagging is performed from left to right and the right context is not yet disambiguated, we need to refer to the ambitags of the tokens on the right hand side. Not only (ambi)tags can be added, but the actual tokens as well. This is usually only useful for highly frequent tokens, such as functors and punctuation marks.

For the unknown-words classifier, contextual information can be used as well (although no ambitag can be constructed for an unknown word). Additionally, some orthographic features are supported by MBT as well, such as the first or last *n* graphemes, or more general features such as hyphenation or capitalization features. While these features work well for unknown words, MBT is mostly driven by contextual features. To contrast it with the approach described in Section 4, we will refer to MBT as a *context-driven tagger*.

Experiments

To obtain reliable accuracy scores for our data sets, we use the technique of ten-fold cross validation to evaluate the respective techniques: the entire corpus is split into ten slices of equal size. Each slice is used as an evaluation set once, while eight slices are used to train the tagger and the remaining slice is used as a validation set to establish the optimal set of features.

The optimal features for part-of-speech tagging cannot be predetermined, as each data set requires different combinations of features. To facilitate the process of feature selection, we automated the search for optimal features: through the stepwise addition of features to the MBT classifiers and the observation of changes in the tagging accuracy on the validations set, we can dynamically establish optimal features for each data set (cf. infra for a discussion on feature selection).

The final evaluation of the tagger is performed on the held-out evaluation set, which is not used at any point during training, thereby establishing accuracy figures on truly unseen data. Table 2 displays the results of the 4x10 experiments. We provide scores for known words, unknown words and the overall tagging accuracy.

The first line in each sub-table displays baseline accuracies, achieved by using a simple unigram tagger, which always selects the most frequent tag for each known word and the overall most frequent tag for unknown words. For Swahili, the baseline accuracy is 97%, which can mostly be attributed to the sheer size and the silver standard nature of the corpus. Baseline tagging results for Northern Sotho (88%), Zulu (86%) and Cilubà (83%) on the other hand, are not so good.

Using MBT improves overall tagging accuracy over the baseline for all data sets under investigation. Swahili can be tagged with a projected accuracy of 98.16% and even unknown words are handled fairly well by this tagger (89.81%). The previously reported result for Northern Sotho (93.5%, cf. de Schryver and De Pauw (2007)) is sig-

	Known Words	**Unknown Words**
Swahili	wddfaaww	chssspppppppwddddFaww
N. Sotho	dddfaaa	csppdFa
Zulu	df	ssspppwdF
Cilubà	wddf	sspppddFa

Table 3: Optimal features for memory-based tagging of Swahili, Northern Sotho, Zulu and Cilubà data sets (majority vote over ten folds).

nificantly improved up to 94.68%, thanks to the additional training data.

The Zulu tagger is able to tag known words almost perfectly, although MBT scores slightly lower than the baseline method. This indicates that tagging known tokens basically amounts to table look-up for this data and that the extra features MBT uses, are not helpful. The underwhelming score for (the copious number of) Zulu unknown words drags the overall tagging accuracy down to 92.71%. A similar situation can be observed for memory-based tagging of Cilubà, although known words accuracy barely exceeds baseline accuracy. This is undoubtedly due to the diminutive size of the corpus.

Feature Selection

Table 3 provides an overview of the optimal features, automatically selected during the development phase, for each of the four data set. Some general tendencies can be observed: for the prediction of unknown words, contextual features (**d** for left context and **a** for ambiguous right context) play a fairly unimportant role, except for Swahili. Particularly the right context does not seem very informative, neither for the prediction of tags for unknown words, nor for known words. Prefix (**p**) and suffix features (**s**) on the other hand are abundantly used for the prediction of tags for unknown words. Capitalization (**c**) and hyphenation (**h**) features are used for the Swahili and Northern Sotho data as well. The use of word tokens (**w**) as an information source is fairly limited, except for the more expansive data sets.

For known words, it can be observed that the disjunctively written language of Northern Sotho benefits heavily from contextual information, as is to be expected. Also for the large Swahili data set, contextual features play an important role. For the more diminutive Zulu and Cilubà data sets however, contextual features are fairly sparsely used during disambiguation.

The experimental results show that context-driven tagging is a viable solution for the languages under investigation, with the exception of Cilubà, which simply does not have enough data available to it to trigger any kind of useful contextual information source. For all of the languages under investigation, the handling of unknown words poses the biggest limitation on achieving state-of-the-art tagging accuracy. In the next section, we will describe an alternative, *resource-light* approach that is able to overcome this bottleneck, while also limiting the observed decrease in known words tagging accuracy.

4. Bag-of-Substrings Tagging

While state-of-the-art tagging accuracy can be achieved for Swahili and Northern Sotho and a fairly reasonable accuracy for Zulu, we need to take into account that the majority of the languages on the African continent are more akin to Cilubà in terms of linguistic resources. Most Sub-Saharan African languages are in fact decidedly resource-scarce: digital linguistic resources are usually not available and while data can be mined off the Internet in a relatively straightforward manner (Hoogeveen and De Pauw, 2011), annotated corpora are few and far between.

In previous research efforts, we have investigated ways to circumvent this by means of projection of annotation (De Pauw et al., 2010; De Pauw et al., 2011) and by means of unsupervised learning techniques (De Pauw et al., 2007; De Pauw and Wagacha, 2007). The latter research efforts attempted to automatically induce morphological features on the basis of a raw, unannotated lexicon of words in the respective target languages. In this section, we will describe how we can use the same technique, dubbed *Bag-of-Substrings*, to perform part-of-speech tagging on the basis of scarce linguistic resources.

The general idea behind the Bag-of-Substrings approach is simple: each token is described as a collection of its substrings, which function as features towards some kind of classification task, in this case part-of-speech tagging. We will illustrate the conversion on the basis of the tagged Swahili Example 2[2]:

(2) Adam$_{PROPNAME}$ alionekana$_{V}$ chumbani$_{N}$
kwake$_{PRON}$ hana$_{NEG}$ fahamu$_{N}$.$_{FULL-STOP}$

This is converted into the representation in Figure 1. For each word we list all of the possible substrings and indicate whether it occurs at the beginning (P=), end (S=) or middle (I=) of the word or whether it constitutes the word itself (W=). These orthographic features encode a lot of potentially useful morphological information, although most features are not relevant towards the actual prediction of the class, i.e. part-of-speech tag.

The advantage of this approach is that we do not need to predefine which features are needed for classification, nor do we require any knowledge about the morphology of the language in question. All of the features are presented to a maximum entropy-based, machine learning classifier (Le, 2004), which will automatically determine during the training phase which of these features are salient in terms of their predictive power. For example, the feature P=A for the word *Adam* implicitly encodes the capitalization of the word and will therefore probably be strongly correlated with the tag propname. Likewise, the features P=a and I=li for the word *alionekana* encode its prefixes. Furthermore, the W= features still enable basic table look-up functionality for known words.

This method of part-of-speech tagging effectively takes all contextual information out of the equation. Instead, all of

[2] *Adam appeared to be distraught.*

Class	Features
PROPNAME	P=A P=Ad P=Ada **W=Adam** I=d I=da S=dam I=a S=am S=m
V	P=a P=al P=ali P=alio P=alion P=alione P=alionek P=alioneka P=alionekan **W=alionekana** I=l I=li I=lio I=lion I=lione I=lionek I=lioneka I=lionekan S=lionekana I=i I=io I=ion I=ione I=ionek I=ioneka I=ionekan S=ionekana I=o I=on I=one I=onek I=oneka I=onekan S=onekana I=n I=ne I=nek I=neka I=nekan S=nekana I=e I=ek I=eka I=ekan S=ekana I=k I=ka I=kan S=kana I=a I=an S=ana I=n S=na S=a
N	P=c P=ch P=chu P=chum P=chumb P=chumba P=chumban **W=chumbani** I=h I=hu I=hum I=humb I=humba I=humban S=humbani I=u I=um I=umb I=umba I=umban S=umbani I=m I=mb I=mba I=mban S=mbani I=b I=ba I=ban S=bani I=a I=an S=ani I=n S=ni S=i
PRON	P=k P=kw P=kwa P=kwak **W=kwake** I=w I=wa I=wak S=wake I=a I=ak S=ake I=k S=ke S=e
NEG	P=h P=ha P=han **W=hana** I=a I=an S=ana I=n S=na S=a
N	P=f P=fa P=fah P=faha P=faham **W=fahamu** I=a I=ah I=aha I=aham S=ahamu I=h I=ha I=ham S=hamu I=a I=am S=amu I=m S=mu S=u
FULL-STOP	**W=.**

Figure 1: Bag-of-Substrings representation of Example 2.

the words are handled by one and the same orthography-based classifier (in contrast to MBT). While this classifier is basically a unigram classifier (cf. Baseline in Table 2) it draws its predictive power from the Bag-of-Substrings (henceforth BoS), presented as training material.

The BoS lines in Table 2 display the experiment results. For Swahili known words, the BoS-approach underperforms, compared to MBT and even the baseline tagger. For unknown words, on the other hand, tagging accuracy is significantly higher using a BoS-approach[3]. The same trend is visible for the Northern Sotho data. For Zulu, baseline accuracy is restored and a huge improvement is achieved in the handling of unknown words, compared to MBT. For the Cilubà data, the BoS approach does not significantly underperform for known words, while it substantially improves for unknown words.

Given the modular design of MBT, we can now envision a mixed tagging approach, using different, individual classifiers for known and unknown words. The results of this virtual experiment are displayed on the last line of Table 2. For Zulu, there is no advantage of using a mixed approach, but for Swahili, Northern Sotho and Cilubà, the optimal combination involves tagging known words with MBT and unknown words with BoS, yielding higher scores than the individual taggers. In practice, tagging accuracy will be even higher for such a combination, since more accurate unknown words tagging will lead to more accurate left-context information for the prediction of known words.

It is important to point out that the BoS approach does not necessarily need a tagged corpus as training material: a simple lexicon in which each word is associated with a part-of-speech tag, is all the training data this approach needs. This is good news, since state-of-the-art tagging accuracies can now be achieved without the development of a manually part-of-speech tagged corpus, unless of course, we are dealing with a language with a disjunctive orthography. As the experimental results for Northern Sotho in Table 2 show, contextual information is essential in such a case and this can only be induced from a tagged corpus. An additional advantage of the BoS approach is its efficiency: Swahili is tagged by MBT at a rate of 4,400 words per second, whereas BoS can tag words at 17,000 words per second.

5. Learning Curves

Table 2 shows that context-driven taggers still have the edge when trained on expansive data sets, whereas the situation is somewhat reversed for the resource-scarce languages of Zulu and Cilubà. In this section, we will perform a direct comparison between the two approaches, using steadily increasing amounts of data. We re-use the slices of the ten-fold cross validation experiments for this. The last slice is kept constant as the evaluation set throughout the experiments. In the first experiment Slice0 (10% of the data) is used as training material. In the next experiment Slice0+Slice1 (20% of the data) is used, etc.

This results in learning curves (Figure 2), which show how the two approaches compare to one another for different data set sizes. For Swahili and Northern Sotho, the learning curves confirm the results of the ten-fold cross validation experiment: MBT consistently outperforms BoS for known words, even for smaller data set sizes, while BoS has the edge for unknown words. For Northern Sotho, a peculiar situation arises in the middle of the experiment: BoS unknown word accuracy is actually higher than BoS known words accuracy. This proves that BoS is not suitable for tagging a language with a disjunctive writing system, even though as an unknown words predictor, it works better than a context-driven tagger does.

For both Zulu and Cilubà, performance for known words is more or less equal for the two techniques. The difference is made in the handling of unknown words. Increasing corpus size for Zulu does not allow MBT to make significantly more accurate predictions and at some points, the addition of new material appears to actually confuse the classifier. BoS is more stable in this respect, as accuracy for tagging Zulu unknown words steadily increases with data set size. For Cilubà BoS tagging of unknown words, on the other

[3] The McNemar Significance test for paired classifiers (McNemar, 1947) was used to establish statistical significance throughout this paper.

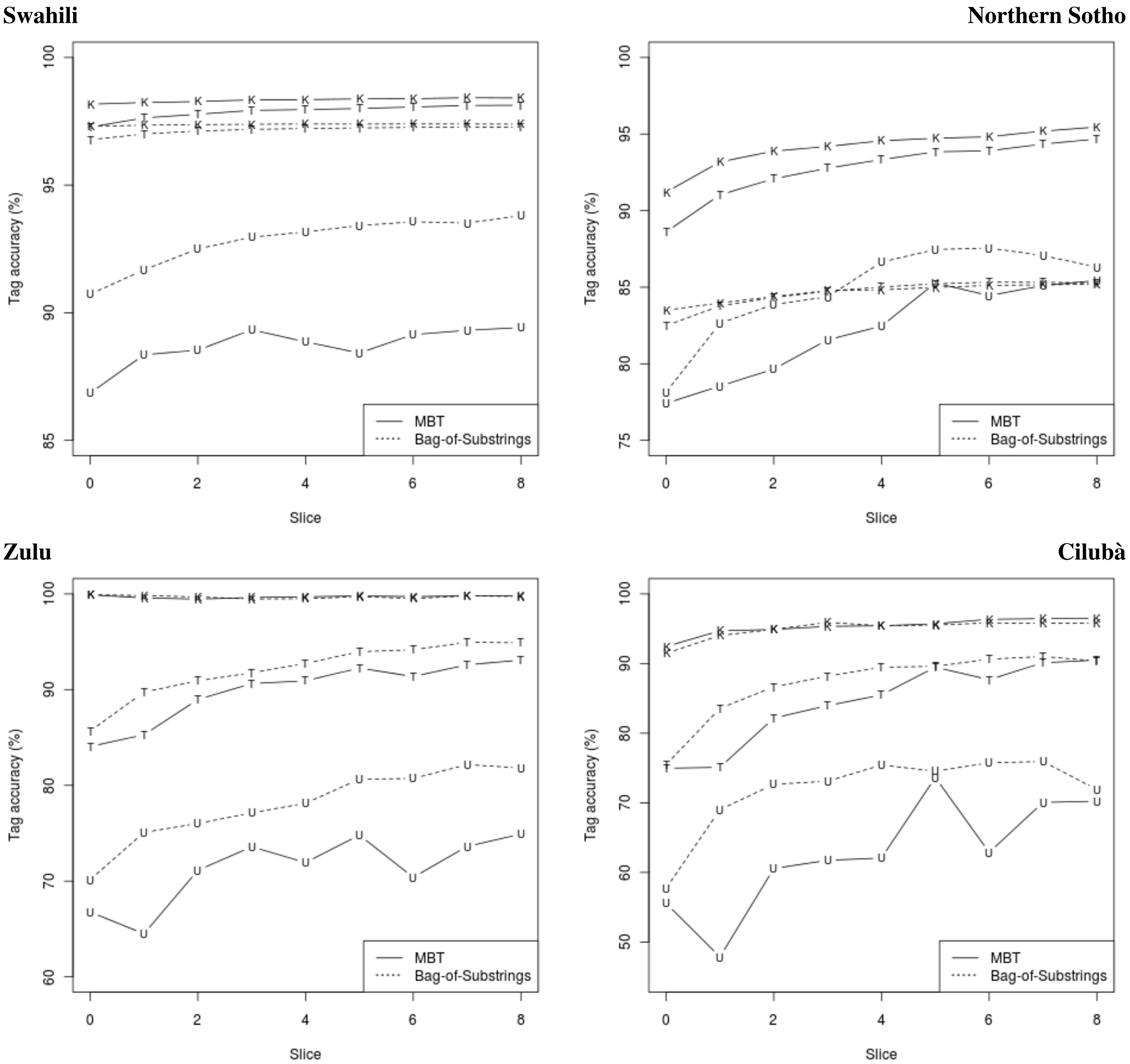

Figure 2: Graphs for learning curve experiments. Learning curves are displayed for (K)nown words, (U)known words and the (T)otal tagging accuracy.

hand, the learning curve seems to plateau and even drop at the end. The erratic curve for MBT however indicates that the small size of the corpus makes the results vulnerable to small distributional changes within the slices and it is not possible to draw any reliable conclusions from this experiment for Cilubà.

6. Conclusion and Future Work

This paper adds to the growing number of publications that describe data-driven approaches to natural language processing of African languages. We described experiments with an off-the-shelf, data-driven tagger and observed reasonable to excellent tagging accuracies for Swahili, Northern Sotho and Zulu. The underwhelming results for Cilubà can be attributed to the diminutive nature of the data set used for training.

The Bag-of-Substrings technique, introduced in this paper as a part-of-speech tagging approach, has been empirically shown to be able to hold its own against a context-driven tagger, provided there is a critical amount of data available. For a language with a disjunctive orthography, the BoS approach is only really useful for the prediction of tags for previously unseen words. The learning curves provided further evidence that the BoS approach establishes a fast, resource-light technique for part-of-speech tagging of agglutinative languages.

In future research efforts, we will also investigate other data-driven taggers. Preliminary experiments with TnT (Brants, 2000) and SVMTool (Giménez and Màrquez, 2004) exhibited the same trends as MBT using the same features, albeit with the former significantly and surprisingly underperforming compared to MBT. This unexpected result in itself warrants further research, which may provide some insight into best practices for Bantu part-of-speech tagging. Since the BoS approach is de facto a data-driven tagger, this experiment can be easily replicated for other languages and data sets. Future research will investigate whether the technique can prove valuable for other Bantu languages and other language groups as well. We will also attempt to introduce contextual features to the maxent classifier that underlies the BoS method, which may serve to get the best of both worlds: accurate unknown word POS-tagging, coupled with context-aware disambiguation of known tokens. Finally, we also aim to further explore the possibility of using lexicons, rather than the much less readily available annotated corpora, as training material for the BoS approach to bootstrap part-of-speech taggers for a wide range of resource-scarce languages.

Acknowledgments and Demos

The first author is funded as a Postdoctoral Fellow of the Research Foundation - Flanders (FWO). We would like to thank Sebastian Dom for his Cilubà annotation efforts and the reviewers for their insightful comments.

Demonstration systems can be found at AfLaT.org for part-of-speech tagging of
Swahili (`http://aflat.org/swatag`),
Northern Sotho (`http://aflat.org/sothotag`),
Zulu (`http://aflat.org/zulutag`) and
Cilubà (`http://aflat.org/lubatag`).

7. References

Brants, T. (2000). TnT a statistical part-of-speech tagger. In *Proceedings of the Sixth Conference on Applied Natural Language Processing (ANLP 2000)*. Seattle, USA: pp. 224–231.

Daelemans, W., Zavrel, J., van den Bosch, A. & Van der Sloot, K. (2010). MBT: Memory-based tagger, version 3.2, reference guide. Technical Report 10-04, University of Tilburg.

De Pauw, G., de Schryver, G-M & Wagacha, P.W. (2006). Data-driven part-of-speech tagging of Kiswahili. In P. Sojka, I. Kopeček & K. Pala (Eds.), *Proceedings of Text, Speech and Dialogue, Ninth International Conference*. volume 4188/2006 of *Lecture Notes in Computer Science*, Berlin, Germany: Springer Verlag, pp. 197–204.

De Pauw, G., Maajabu, N.J.A. & Wagacha, P.W. (2010). A knowledge-light approach to Luo machine translation and part-of-speech tagging. In G. De Pauw, H. Groenewald & G-M de Schryver (Eds.), *Proceedings of the Second Workshop on African Language Technology (AfLaT 2010)*. Valletta, Malta: European Language Resources Association (ELRA), pp. 15–20.

De Pauw, G. & Wagacha, P.W. (2007). Bootstrapping morphological analysis of Gĩkũyũ using unsupervised maximum entropy learning. In *Conference Program and Abstract Book of the Eighth Annual Conference of the International Speech Communication Association*. Antwerp, Belgium, 29 August: ISCA, p. 119.

De Pauw, G., Wagacha, P.W. & Abade, D.A. (2007). Unsupervised induction of Dholuo word classes using maximum entropy learning. In K. Getao & E. Omwenga (Eds.), *Proceedings of the First International Computer Science and ICT Conference*. Nairobi, Kenya: University of Nairobi, pp. 139–143.

De Pauw, G., Wagacha, P.W. & de Schryver, G-M. (2011). Exploring the SAWA corpus - collection and deployment of a parallel corpus English - Swahili. *Language Resources and Evaluation - Special Issue on African Language Technology*, 45(3), pp. 331–344.

de Schryver, G-M & De Pauw, G. (2007). Dictionary writing system (DWS) + corpus query package (CQP): The case of TshwaneLex. *Lexikos*, 17, pp. 226–246.

Dione, C.M.B., Kuhn, J. & Zarrie, S. (2010). Design and development of part-of-speech-tagging resources for Wolof (Niger-Congo, spoken in Senegal). In N. Calzolari, Kh. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner & D. Tapias (Eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA).

Gambäck, B., Olsson, F., Argaw, A.A. & Asker, L. (2009). Methods for Amharic part-of-speech tagging. In G. De Pauw, G-M de Schryver & L. Levin (Eds.), *Proceedings of the First Workshop on Language Technologies for African Languages (AfLaT 2009)*. Athens, Greece: Association for Computational Linguistics, pp. 104–111.

Giménez, J. & Màrquez, L. (2004). A general POS tagger generator based on support vector machines. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*. Lisbon, Portugal: pp. 43–46.

Hahn, S., Lehnen, P., Raymond, C. & Ney, H. (2008). A comparison of various methods for concept tagging for spoken language understanding. In N. Calzolari, Kh. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis & D. Tapias (Eds.), *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco: European Language Resources Association (ELRA).

Hoogeveen, D. & De Pauw, G. (2011). Corpuscollie - a web corpus mining tool for resource-scarce languages. In *Proceedings of Conference on Human Language Technology for Development*. Alexandria, Egypt: Bibliotheca Alexandrina, pp. 44–49.

Hurskainen, A. (2004a). HCS 2004 – Helsinki Corpus of Swahili. Technical report, Compilers: Institute for Asian and African Studies (University of Helsinki) and CSC.

Hurskainen, A. (2004b). Swahili language manager: A storehouse for developing multiple computational applications. *Nordic Journal of African Studies*, pp. 363–397.

Le, Z. (2004). Maximum entropy modeling toolkit for python and c++. Technical report, Available at: http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html. (Accessed: 2 March 2012).

Marcus, M., Santorini, B. & Marcinkiewicz, M. (1993). Building a large annotated corpus of English: The Penn treebank. *Computational Linguistics*, 19(2), pp. 313–330.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2), pp. 153–157.

Prinsloo, D.J. & de Schryver, G-M. (2002). Towards an 11 x 11 array for the degree of conjunctivism / disjunctivism of the South African languages. *Nordic Journal of African Studies*, 11(2), pp. 249–265.

Ramshaw, L.A. & Marcus, M.P. (1995). Text chunking using transformation-based learning. In D. Yarowsky & K. Church (Eds.), *Proceedings of the Third ACL Workshop on Very Large Corpora*. Cambridge, USA: Association for Computational Linguistics, pp. 82–94.

Ratnaparkhi, A. (1996). A maximum entropy model for part-of-speech tagging. In E. Brill & K. Church (Eds.), *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Philadelphia, USA: Association for Computational Linguistics, pp. 133–142.

Spiegler, S., van der Spuy, A. & Flach, P.A. (2010). Ukwabelana - an open-source morphological Zulu corpus. In Q. Lu & T. Zhao (Eds.), *Proceedings of the 23rd International Conference on Computational Linguistics*. Beijing, China: Tsinghua University Press, pp. 1020–1028.

Veenstra, J., van den Bosch, A., Buchholz, S., Daelemans, W. & Zavrel, J. (1999). Memory-based word sense disambiguation. In F. Van Eynde (Ed.), *Computational linguistics in the Netherlands 1998*, pp. 81–92: Rodopi, Amsterdam.