A Memory-Based Approach to Kĩkamba Named Entity Recognition

Benson N. Kituku[1] **Peter W. Wagacha**[1] **Guy De Pauw**[2]
[1]School of Computing & Informatics [2]CLiPS - Computational Linguistics Group
University of Nairobi University of Antwerp
Nairobi, Kenya Antwerp, Belgium
nebsonkituku@yahoo.com guy.depauw@ua.ac.be
waiganjo@uonbi.ac.ke

Abstract

This paper describes the development of a data-driven part-of-speech tagger and named entity recognizer for the resource-scarce Bantu language of Kĩkamba. A small web-mined corpus for Kĩkamba was manually annotated for both classification tasks and used as training material for a memory-based tagger. The encouraging experimental results show that basic language technology tools can be developed using limit amounts of data and state-of-the-art language-independent machine learning techniques.

1 Introduction

The issue of Named Entity Recognition was one of the four themes of the Sixth Message Understanding Conference (MUC-6) (Grishman and Sundheim, 1996). Although the focus was on defense related articles then, there has been a tremendous increase in research efforts over the years for different domains and languages, as presented in Nadeau and Sekine (2007). Named entity recognition (henceforth *NER*) can be defined as the task of recognizing specific concepts in a text, such as proper names, organizations, locations and the like. Part-of-speech tagging (POS tagging) is often mentioned in the same breath as NER, as it is used as an essential pre-processing step to accurate NER. POS tagging can be defined as assigning morphosyntactic categories to words.

Kĩkamba (Kamba) is a Bantu language spoken by almost four million Kamba people in Kenya, according to the 2009 population & housing census (Oparany, 2010). Most of this population lives in the Machakos and Kitui counties and a substantial number along the Embu, Taita Taveta and Tharaka boundaries. For a long time the Kamba people have preserved their culture through carving, especially at Wamunyu and also basketry (*kĩondo*) and traditional dance (*kĩlumi*). The Akamba Culture Trust (ACT) formed in 2005, is crusading for the preservation of culture through written form in literature and research departments. Despite the efforts of the organization and the number of people speaking the language, Kĩkamba still lacks basic language technology resources and tools. Only recently a spell checker was developed at the School of Computing & Informatics of the University of Nairobi in Kenya.

This paper focuses on the development of a Named Entity Recognizer for Kĩkamba. Having a good NER system for this language is useful for a wide range of applications, such as event detection with an emphasis on map and phrase browsing, information retrieval and general data mining. Building a successful NER system cannot really be done without an accurate part-of-speech tagger, unfortunately not available for Kĩkamba. In this paper we will outline how a part-of-speech tagger and named-entity recognizer can be built with a minimum of human effort, using annotated corpora and language-independent, state-of-the-art machine learning techniques.

2 Related Research

A wide variety of languages have been examined in the context of named entity recognition (Nadeau and Sekine, 2007) and part-of-speech tagging, but very few sub-Saharan African languages have such

Conference on Human Language Technology for Development, Alexandria, Egypt, 2-5 May 2011.

tools available to them. Part-of-speech Tagging has been investigated in the South African language context. A number of tag sets and preliminary systems are available for Setswana (van Rooy and Pretorius, 2003), Xhosa (Allwood et al., 2003), Northern Sotho (Prinsloo and Heid, 2005; Taljard and Bosch, 2005; de Schryver and De Pauw, 2007; Faaß, 2010). Outside of South Africa, POS tagging for Swahili has been extensively researched using finite-state-techniques (Hurskainen, 2004) and machine learning methods (De Pauw et al., 2006) and some preliminary experiments on Luo have been described in De Pauw et al. (2010).

Swahili is also - to the best of our knowledge - the only Bantu language that has been studied in the context of named-entity recognition (Shah et al., 2010). A few research efforts however investigate the problem of recognizing African named entities in regional varieties of English, such as South African English (newspaper articles) (Louis et al., 2006) and Ugandan English (legal texts) (Kitoogo et al., 2008).

3 Approaches in building the classifier

There are roughly two design options available when building a classifier for NER. The first one involves hand crafting a dictionary of names (*a gazetteer*) and an extensive list of hand-written disambiguation rules. This option is time consuming, particularly for a less-studied language such as Kĩkamba. Another option is to use techniques that learn the classification task from annotated data. This has the advantage that different techniques can be investigated for the same annotated corpus and that evaluation is possible by comparing the output of the classifier to that of the reference annotation (see Section 6).

As our machine learning algorithm of choice, we have opted for Memory-Based Learning (MBL) (Daelemans and van den Bosch, 2005). MBL is a lazy learning algorithm that simply takes the training data and stores it in memory. New data can be classified by comparing it to the items in memory and extrapolating the classification of the most similar item in the training data. For our experiments we used MBT (Daelemans et al., 2011a), which is a wrapper around the memory-based learning software TiMBL (Daelemans et al., 2011b) that facilitates the learning of sequential classficiation tasks.

4 Corpus Annotation

To build a machine-learning classifier for POS tagging and NER, an annotated corpus needs to be built. Previous research (de Schryver and De Pauw, 2007) showed that it is possible to quickly build a POS tagger from scratch on the basis of a fairly limited amount of data. The experiments described in this paper explore this train of thought for the Kĩkamba language and further extend it to the NER classification task.

A manually cleaned and automatically tokenized web-mined corpus of about 28,000 words was manually annotated for parts-of-speech and named entities. For the former annotation task a very small tag set was used that identifies the following parts of speech: `noun`, `verb`, `adverb`, `adjective`, `preposition`, `punctuation`, `interjection`, `cardinal`, `pronoun` and `conjunction`. These coarse-grained tags can be used in future annotation efforts as the basis for a more fine-grained tag set.

The NER annotation uses the IOB tagging scheme, originally coined by Ramshaw and Marcus (1995) for the natural language processing task of phrase chunking. The IOB scheme indicates whether a word is inside a particular named entity (`I`), at the beginning of an entity (`B`)[1] or outside of an entity (`O`). We distinguish between three types of named entities, namely persons (`PER`), organizations (`ORG`) and locations (`LOC`).

Since one of the main bottlenecks of NER for non-Indo-European languages is the lack of gazetteers of foreign names, we also added an additional 2000 Kĩkamba words for place, people and organization names. These were also used to facilitate and speed up the annotation process.

Manual annotation of the words was done using a spreadsheet. This manual process also helped to detect anomalies (and errors) which had not been resolved during the cleaning stage, hence improving the quality of the the classifier. Each word was placed on separate row (token) with subsequent

[1] In practice, the `B` tag is only used to mark the boundary between two immediately adjacent named entities and is therefore relatively rare.

Token	POS Tag	NER category
Ũsumbĩ	noun	I-ORG
wa	preposition	O
Ngai	noun	I-PER
nĩ	conjunction	O
kyaũ	adjective	O
?	punc	O

Table 1: Sample annotation of Kĩkamba corpus.

columns providing a drop-down box for parts of speech and named entity classes. A very small sample of the annotated corpus can be found in Table 1.

5 Features for Classification

During classification words are handled differently according to whether they are considered to be *known* or *unknown*. Known words are tokens that have been encountered in the training data and for which classification can usually be accurately done on the basis of local context by looking at the surrounding words and classes. For unknown words, i.e. words that have never been seen before, we also add pseudo-morphological information to the information source, such as the first and last *n* characters of the word to be tagged and information about hyphens and capitalization. Particularly the last feature is important, since during POS tagging the identification of (proper) nouns is paramount to the NER system.

The Memory-based Tagger (MBT) builds two separate classifiers for known and unknown words. The optimal set of features for classification was experimentally established. For known words, the best context considered two disambiguated part-of-speech tags to the left of the word to be tagged and one (not yet disambiguated) tag to the right. The accuracy of the part-of-speech tagger (> 90%) can be found in Table 3.

For the unknown words group we included the aforementioned pseudo-morphological features alongside the typical contextual ones. We used a local disambiguation context of only one tag on both sides. Increasing the context size to be considered resulted in an increase in classification errors: the training data is apparently too limited to warrant a larger context to generalize from when it comes to classifying unknown words. The average tagging accuracy of 71.93% (Table 3) shows that more data will be needed to arrive at a more accurate handling of unknown words.

In view of the morphological structure of the Kĩkamba language many words will start with Mb (e.g. Mbui - goat), Nd (Ndua - village), Ng (Ng'ombe - cow), Ny (Nyamu - wild animal), Th (Thoaa - price), Mw (Mwaka - year), Kw (Kwangolya - place name), Ky (Kyeni - light), Ma (maiu - bananas), Sy (Syombua - person name). These examples show that even considering only the first two letters of unknown words is a good idea, as these are usually quite indicative of their morphosyntactic class, in this case the nominal class.

Furthermore, we also consider the last two letters, as most names of places in the Kĩkamba language will end in *-ni*, e.g. Kathiani, Kaviani, Nzaikoni, Makueni and Mitaboni. As mentioned before we also include capitalization as an import feature towards disambiguation. The hyphenation feature however did not provide much improvement in terms of accuracy. For the Kĩkamba language an interesting feature would indicate the presence of a single-quote (') in the word, as this can also be an informative discriminating factor (e.g. for the words Ng'aa, Ng'ala, Ng'anga, Ng'eng'eta, Ng'ombe, Ng'ota etc.). In future work, we will investigate ways to introduce such language-specific orthographic features in the machine learning approach.

6 Experiments and Results

In this section we will describe the experimental results obtained on the basis of our small annotated corpus of 28,000 words. Various metrics were used for the evaluation of both the POS tagger and the NER system: accuracy, precision, recall and F-score. Accuracy simply expresses the number of times the classifier made the correct classification decision. Precision on the other hand calculates for each class how many times the class was correctly predicted, divided by the number of times that particular class was predicted in total. Recall on the other hand is defined as the number of times a class was correctly predicted, divided by the number of

Metric	POS tagging	NER
Precision	83.24	96.47
Recall	72.34	87.13
F-score	77.41	91.56

Table 2: Recall, precision and F-score for both classifiers

times that particular class appears in the test data. Finally, the F-score is the harmonic mean of recall and precision, calculated as outlined in Formula 1. The precision weighting factor β was set to 1.

$$F_\beta = (1 + \beta^2).\frac{precision.recall}{\beta^2.precision + recall} \quad (1)$$

6.1 Experimental Setup

The k-folds evaluation method (Weiss and Kulikowsie, 1991) was used to perform the evaluation of the system. This method was selected because of the relatively small size of the Kĩkamba corpus. We used a *k* value of 10. This mean the annotated corpus is randomly portioned in ten equal folds (respecting sentence boundaries). For each experimental *fold*, one part was used as the evaluation set, while the other nine parts made up the training set. This ensures that evaluation takes place over all of the annotated data. The biggest advantage of this kind experimental setup is that we can measure the accuracy on data that is not known to the system during training. By comparing the output of the classifiers to the annotation, we can calculate the aforementioned evaluation scores.

6.2 Results

Tables 2 and 3 outline the experimental results. The latter shows the accuracy for both classification tasks and for each of the ten partitions, followed by the average score. The former table shows precision, recall and F-score. These figures were obtained by calculating precision and recall for each class (i.e. each part-of-speech tag and named entity class) and averaging the scores.

Table 2 shows a precision of 83.24% and 96.47% for POS tagging and NER respectively. Error analysis showed that for the part of speech tagging task there was substantial false positive classification which lowered the percentage. A closer look at the individual class categories for POS tagging and the confusion matrix extracted from MBT, indicates that the noun and preposition classes were particularly vulnerable to the effect of false positives. For the NER system the least false negatives were seen for class "O" with the other classes doing reasonably well too.

The recall scores in Table 2 are significantly lower than the precision scores. The low recall score for part-of-speech tagging is mainly due to a rather bad handling of verbs and numerals. The latter result is surprising since numerals should be straightforward to classify. More worryingly is the bad recall score for verbs. Future work will include an error analysis to identify the bottleneck in this case. The recall scores for NER on the other hand are more encouraging. The main bottleneck here is the handling of `B-` type tags. Most likely there is not enough data to handle these rather rare classes effectively.

Finally, the F-score stands at 77.41% and 91.56% for POS tagging and NER respectively. The F-score for POS tagging suffers because of the recall problem for verbs, but the F-score for NER is very encouraging, also compared to the results of Shah et al. (2010), who describe an alternative approach to NER for the Bantu language of Swahili.

Table 3 includes separate scores for known and unknown words. A closer look at the results for POS tagging indicates that, particularly given the very limited size of the training data, known words are actually handled pretty accurately (94.65%). Unknown words fare a lot worse at 71.93% accuracy. Error analysis shows that this is related to the aforementioned problem of verbal classification. A more well-rounded approach to modeling morphology within the classifier could provide a significant increase in unknown word tagging accuracy.

Compared to the results of the Swahili part-of-speech tagger described in De Pauw et al. (2006), the Kĩkamba system still has a long way to go. The Swahili tagger scored 98.46% and 91.61% for known and unknown words respectively with an overall performance of 98.25%. The Kĩkamba has an overall accuracy of 90.68%. This is obviously due to the difference in data set size: the Swahili corpus counted more than 3 million tokens, compared to only 28,000 words for the Kĩkamba tagger. Given this restriction, the performance of the tagger is sur-

FOLD	Part-of-Speech Tagging			Named Entity Recognition		
	Known	**Unknown**	**Overall**	**Known**	**Unknown**	**Overall**
1	94.24	78.01	92.07	98.81	98.44	98.76
2	94.59	73.65	90.64	98.13	88.07	96.21
3	94.55	71.31	90.22	98.60	93.00	97.56
4	94.79	68.44	90.75	98.67	94.68	98.07
5	93.44	71.43	89.74	98.77	91.28	97.48
6	94.34	68.62	90.41	98.76	97.56	98.58
7	95.85	67.05	90.43	99.33	95.73	98.64
8	95.46	70.25	91.06	98.81	95.96	98.31
9	95.42	83.64	92.69	98.59	90.63	96.75
10	93.80	66.86	88.74	99.13	95.39	98.42
Av.	94.65	71.93	90.68	98.76	94.07	97.88

Table 3: Experimental Results for the Part-of-Speech Tagging and Named Entity Recognition Tasks (10-fold cross-validation)

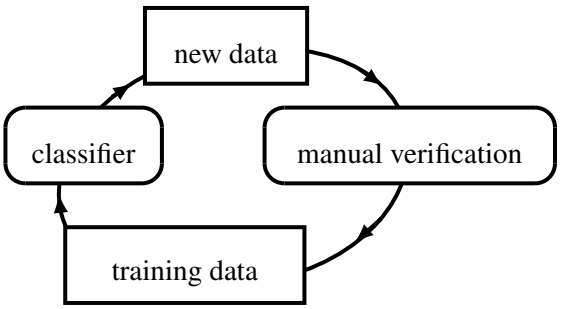

Figure 1: Semi-Automatic Annotation.

prisingly high.

For the NER task we report a performance of 98.76% and 94.07% for known and unknown words respectively, with an overall performance of 97.88%. Again, given the size of the data, this is an encouraging result and further evidence that data-driven approaches, i.e. techniques that learn a classification task on the basis of manually annotated data, are a viable way to unlock the language technology potentials of Bantu languages.

7 Conclusion and Future Work

We have presented a Kĩkamba Named Entity Recognizer with a classification accuracy of 97.88% and an F-score of 91.71% and a part-of-speech tagger with an accuracy of 90.68%. While the amount of data is rather limited and we have not yet performed a full exploration of all experimental parameters, these scores are encouraging and further underline the viability of the data-driven paradigm in the context of African language technology.

We will investigate other machine learning techniques for these classification tasks and this data. As soon as a critical mass of training data is available, we will also perform *learning curve* experiments to determine how much data is needed to arrive at accuracy scores comparable to the state-of-the-art in NER and POS tagging.

At this point, we can use the systems described in this paper, to semi-automatically annotate larger quantities of data. This process is illustrated in Figure 1: we use the currently available training data to train a classifier that automatically annotates new data. This is then checked manually and corrected where necessary. The resulting data can then be added to the training data and a new classifier is trained, after which the cycle continues. This type of semi-automatic annotation significantly improves the speed and consistency with which data can be annotated. As such the systems described in this paper should be considered as the first bootstrap towards an expansive annotated corpus for Kĩkamba.

References

J. Allwood, L. Grönqvist, and A. P. Hendrikse. 2003. Developing a tagset and tagger for the African lan-

guages of South Africa with special reference to Xhosa. *Southern African Linguistics and Applied Language Studies*, 21(4):223–237.

W. Daelemans and A. van den Bosch. 2005. *Memory-based Language Processing. Studies in Natural Language Processing*. Cambridge University Press, Cambridge, UK.

W. Daelemans, J. Zavrel, A. van den Bosch, and K. van der Sloot. 2011a. MBT: Memory Based Tagger, version 3.2, Reference Guide. ILK Research Group Technical Report Series 10-04, Tilburg.

W. Daelemans, J. Zavrel, K. van der Sloot, and A. van den Bosch. 2011b. TiMBL: Tilburg Memory Based Learner, version 6.3, Reference Guide. ILK Research Group Technical Report Series no. 10-01 04-02, Tilburg University.

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

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

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

G. Faaß. 2010. The verbal phrase of Northern Sotho: A morpho-syntactic perspective. In G. De Pauw, H.J. Groenewald, and G-M. de Schryver, editors, *Proceedings of the Second Workshop on African Language Technology (AfLaT 2010)*, pages 37–42, Valletta, Malta. European Language Resources Association (ELRA).

R. Grishman and B. Sundheim. 1996. Message understanding conference-6: a brief history. In *Proceedings of the 16th conference on Computational linguistics - Volume 1*, COLING '96, pages 466–471, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

F. Kitoogo, V. Baryamureeba, and G. De Pauw. 2008. Towards domain independent named entity recognition. In *Strengthening the Role of ICT in Development*, pages 38–49, Kampala, Uganda. Fountain Publishers.

A. Louis, A. De Waal, and C. Venter. 2006. Named entity recognition in a South African context. In *Proceedings of the 2006 annual research conference of the South African institute of computer scientists and information technologists on IT research in developing countries*, SAICSIT '06, pages 170–179, Republic of South Africa. South African Institute for Computer Scientists and Information Technologists.

D. Nadeau and S. Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, January.

W.A. Oparany. 2010. *2009 Population & Housing Census Results*. Available from [http://www.knbs.or.ke/Census Results/Presentation by Minister for Planning revised.pdf], Nairobi, Kenya.

D. J. Prinsloo and U. Heid. 2005. Creating word class tagged corpora for Northern Sotho by linguistically informed bootstrapping. In *Proceedings of the Conference on Lesser Used Languages & Computer Linguistics (LULCL-2005)*, Bozen/Bolzano, Italy.

L. A. Ramshaw and M. P. Marcus. 1995. Text chunking using transformation-based learning. In *Proceedings of the Third Annual Workshop on Very Large Corpora*, pages 82–94. ACL.

R. Shah, B. Lin, A. Gershman, and R. Frederking. 2010. Synergy: A named entity recognition system for resource-scarce languages such as Swahili using online machine translation. In G. De Pauw, H.J. Groenewald, and G-M de Schryver, editors, *Proceedings of the Second Workshop on African Language Technology (AfLaT 2010)*, pages 21–26, Valletta, Malta. European Language Resources Association (ELRA).

E. Taljard and S. E. Bosch. 2005. A comparison of approaches towards word class tagging: disjunctively vs conjunctively written Bantu languages. In *Proceedings of the Conference on Lesser Used Languages & Computer Linguistics (LULCL-2005)*, Bozen/Bolzano, Italy.

B. van Rooy and R. Pretorius. 2003. A word-class tagset for Setswana. *Southern African Linguistics and Applied Language Studies*, 21(4):203–222.

S.M. Weiss and C.A. Kulikowsie. 1991. *Computer systems that learn*. Morgan Kaufmann, San Mateo, CA, USA.