Towards English - Swahili Machine Translation

Guy De Pauw[1,2], **Peter Waiganjo Wagacha**[2], **Gilles-Maurice de Schryver**[3]

[1]CLiPS - Computational Linguistics Group, University of Antwerp, Belgium, guy.depauw@ua.ac.be
[2]School of Computing & Informatics, University of Nairobi, Kenya, waiganjo@uonbi.ac.ke
[3]Department of African Languages and Cultures, Ghent University, Belgium, gillesmaurice.deschryver@ugent.be

The SAWA corpus

Even though the Bantu language of Swahili is spoken by more than fifty million people in East and Central Africa, it is surprisingly resource-scarce from a language technological point of view, an unfortunate situation that holds for most, if not all languages on the continent. The increasing amount of digitally available, vernacular data has prompted researchers to investigate the applicability of corpus-based approaches to *African language technology*. In this vein, the SAWA corpus project attempts to collect and deploy a parallel corpus English - Swahili, not only for the straightforward purpose of developing a machine translation system, but also to investigate the possibility of projection of annotation into a resource-scarce, African language.

Compiling a balanced and expansive parallel corpus English - Swahili is a rather daunting task. While monolingual Swahili data is abundantly available on the Internet, sourcing parallel texts is cumbersome. Even countries that have both English and Swahili as their official languages, such as Tanzania, Kenya and Uganda, do not tend to translate and/or publish all government documents bilingually. One therefore opportunistically collects whatever can be found in the public domain.

At this point in the data collection phase, that means that the 2.2 million word parallel corpus is biased towards religious material, such as bible and quran translations. Nevertheless, the more interesting, secular part of the SAWA corpus ($\pm$ 420k words) is steadily increasing, thanks to the inclusion of bilingual investment reports, manually translated movie subtitles, political documents and material kindly donated by local translators to the SAWA project.

Each text in the SAWA corpus is automatically part-of-speech tagged and lemmatized, using the TreeTagger for the English part (Schmid, 1994) and the systems described in De Pauw et al. (2006) and De Pauw and de Schryver (2008) for Swahili. These extra annotation layers allow us to perform more accurate automatic word alignment on the basis of factored data.

Table 1: Precision, Recall and F-score for the word-alignment task using GIZA++.

	Prec.	**Recall**	**F($\beta = 1$)**
Words	39.4%	44.5%	41.8%
Morphemes	50.2%	64.5%	55.8%
Morphemes + dict	66.5%	72.6%	69.4%

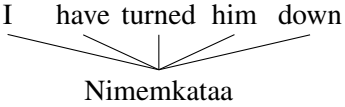
About half of the SAWA corpus was manually sentence-aligned. This part of the corpus was then used to train a supervised sentence alignment approach, based on maximum entropy learning, enabling us to automatically sentence-align the other parts of the SAWA corpus with an estimated accuracy of about 98%. We also compiled a very small, manually word-aligned evaluation set of about 5,000 words. This allows us to perform a limited quantitative evaluation of the automatic word alignment approaches.

Automatic Word Alignment

We performed word-alignment experiments using GIZA++ (Och and Ney, 2003) on the factored data of the SAWA corpus. While this method is well suited to handle closely related languages, the scores for word alignment on the SAWA corpus, even on the basis of factored data, are rather underwhelming (first row in Table 1).

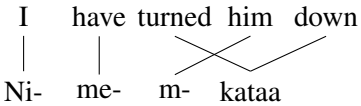
The main problem in training a GIZA++ model for the language pair English - Swahili is the strong agglutinating nature of the latter. No parallel corpus is exhaustive enough to provide enough linguistic evidence to unearth strongly converging alignment patterns, such as the one in Example 1.

(1)



Morphologically deconstructing the word however can greatly relieve the sparse data problem for this task:

(2)



The isolated Swahili morphemes can more easily be linked to their English counterparts, since there will be more linguistic evidence in the parallel corpus, linking for example *ni* to *I* and *m* to *him*. This kind of morphological segmentation can be done in a relatively straightforward manner by post-processing the output of the Swahili lemmatizer.

We have no morphologically aligned gold standard data available, so evaluation of the morpheme-based approach needs to be done in a roundabout way. We first morphologically decompose the Swahili data and run GIZA++ again. Next we recompile the Swahili words from the morphemes and group the word-alignment links accordingly. Incompatible linkages are removed and simple majority voting resolves ambiguous alignment patterns. The updated scores are presented in the second row of Table 1 and show that this type of processing is highly beneficial. Adding a consolidated database of four electronic English - Swahili translation dictionaries (De Pauw et al., 2009a) further improves on the word-alignment scores (third row Table 1).

Machine Translation

The most straightforward and practical application of a parallel corpus is undoubtedly as a resource to build a statistical machine translation (SMT) system. We used the standard MOSES package (Koehn et al., 2007) to construct a machine translation system on the basis of the SAWA corpus. We did not perform extensive parameter tweaking on either the SMT or language model side, mostly restricting ourselves to the default settings. Therefore the experimental results presented in this section leave considerable room for improvement.

The SAWA corpus was randomly divided into a 90% training set and a 10% test set. The SMT system was built on the training set and evaluated on the test set, using the standard machine translation evaluation measures BLEU and NIST. We compare our results to that of the Google Translate system for Swahili. This comparison is problematic: the Google Translate system is partially built on the basis of data described in a previous publication of the SAWA corpus (De Pauw et al., 2009b) and it is obvious that significant portions of the test set in our experiments actually constitute training data for the Google Translate system.

The experimental results can be found in Table 2. For English→Swahili translation the SAWA system underperforms compared to Google Translate's system. This may be partly attributed to the aforementioned evaluation problem, but is also likely due to Google's more refined morphological generation model on the target language side. Error analysis shows that the SAWA system has significant difficulties generating morphologically correct Swahili words.

Table 2: BLEU and NIST scores for Bidirectional Machine Translation Task.

		BLEU	NIST
GOOGLE	**English →Swahili**	0.26	3.96
SAWA	**English →Swahili**	0.20	2.92
GOOGLE	**Swahili →English**	0.29	4.14
SAWA	**Swahili →English**	0.35	4.52

For Swahili→English translation, our system fares better, not hampered by the morphological generation issues of the target language. In this case, the SAWA system is able to outperform the Google system by a significant margin.

Discussion

This abstract presented the development of a parallel corpus English - Swahili and early experiments in machine translation for this language pair. The current version of the SAWA corpus has more than two million words and is part-of-speech tagged, lemmatized and sentence and word-aligned. To our knowledge, this is the only such resource available for a sub-Saharan African language. While the experimental results are modest at this point, we are confident that the inclusion of a more refined Swahili morphological generation component, a thorough parameter exploration of MOSES, as well as a more balanced constitution of future versions of the SAWA corpus will provide significant advances towards more accurate machine translation for this language pair.

Acknowledgments, Demo and Data

The first author is funded as a Postdoctoral Fellow of the Research Foundation - Flanders (FWO). A demonstration machine translation system and parts of the SAWA corpus will be made publicly available through AfLaT.org.

References

De Pauw, G. & de Schryver, G.-M. (2008). Improving the computational morphological analysis of a Swahili corpus for lexicographic purposes. *Lexikos*, 18, pp. 303–318.

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, 9th International Conference*. Berlin, Germany: Springer Verlag, pp. 197–204.

De Pauw, G., de Schryver, G.-M. & Wagacha, Peter Waignjo. (2009)a. A corpus-based survey of four electronic Swahili–English bilingual dictionaries. *Lexikos*, 19, p. 340–352.

De Pauw, G., Wagacha, P.W. & de Schryver, G.-M. (2009)b. The SAWA corpus: a parallel corpus English - Swahili. 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. 9–16.

Koehn, Ph., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A. & Herbst, E. (2007). MOSES: Open source toolkit for statistical machine translation. In *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session*. Prague, Czech Republic: Association for Computational Linguistics, pp. 177–180.

Och, F.J. & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1), pp. 19–51.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In D. Jones (Ed.), *Proceedings of the International Conference on New Methods in Language Processing*. Manchester, UK: UMIST, pp. 44–49.