4

Bootstrapping Machine Translation for the Language Pair English – Kiswahili

Guy De Pauw, Peter Waiganjo Wagacha and Gilles-Maurice de Schryver

In recent years, research in Machine Translation has greatly benefited from the increasing availability of parallel corpora. Processing the same text in two different languages yields useful information on how words and phrases are translated from a source language into a target language. To investigate this, a parallel corpus is typically aligned by linking linguistic tokens in the source language to the corresponding units in the target language. An aligned parallel corpus therefore facilitates the automatic development of a machine translation system. In this paper, we describe data collection and annotation efforts and preliminary experiments with a parallel corpus English - Kiswahili.

1. Introduction

Language technology applications such as speech recognition and machine translation can provide an invaluable impetus in bridging the digital divide. For a language like Kiswahili, digital resources have become increasingly important in everyday life both in urban and rural areas, particularly thanks to the increasing number of web-enabled mobile phone users in the language area. Most research in the field of language technology for African languages is however still firmly rooted in the knowledge-based paradigm, in which language applications and tools are built on the basis of manually compiled rules. This approach makes development of language technology applications expensive, since it requires a significant amount of expert knowledge and manual effort. The need for a cheaper and faster alternative for developing African language technology applications is therefore high.

The data-driven, corpus-based approach envisioned in this paper establishes such an alternative, so far not yet extensively investigated for African languages. The main advantage of this approach is its language independence: all that is needed is (linguistically annotated) language data, which is fairly cheap to compile. Given this data, existing state-of-the-art algorithms and resources can consequently be re-used to quickly develop robust language applications and tools.

Most African languages are however resource-scarce, meaning that digital text resources are few. An increasing number of publications however are showing that carefully selected procedures can indeed bootstrap language technology for Kiswahili [De Pauw et al. 2006] and even smaller local Kenyan languages [De Pauw and Wagacha 2007; De Pauw et al. 2007a; De Pauw et al. 2007b].

In this paper we outline on-going research on the development of a data-driven machine translation system for the language pair English – Kiswahili. We first provide a short survey of the different approaches to machine translation (Section 2). We then concentrate on the required data collection and annotation efforts (Section 3) and describe some preliminary experimental results with automatic sentence and word alignment tools (Section 4). We conclude with a discussion of the current limitations to the approach and provide pointers for future research.

2. Machine Translation

The main task of Machine Translation (MT) can be defined as having a computer take a text input in one language, the Source language (SL), decode its meaning and re-encode it producing as output a similar-meaning text in another language, the Target language (TL). The idea of building an application to automatically convert text from one language to an equivalent text-meaning in a second language traces its roots back to cold ward intelligence efforts in the 1950's and 60's for Russian-English text translations. Since then a large number of MT systems have been developed with varying degrees of success. For an excellent overview of the history of MT, we refer the reader to [Hutchins 1986].

The original dream of creating a fully automatic MT system has long since been abandoned and most research in the field currently concentrates on minimizing human pre- and post-processing effort. A human translator is thus considered to work alongside the MT system to produce faster and more consistent translations.

The Internet brought in an interesting new dimension to the purpose of MT. In the mid 1990s, free online translation services began to surface with an increasing number of MT vendors. The most famous example is AltaVista's Babelfish, offering on-line versions of Systran to translate English, French, German, Spanish and other Indo-European languages. Currently Google.inc is also offering translation services. While these systems provide far from perfect output, they can often give readers a sense of what is being talked about on a web page in a language (and often even character set) foreign to them.

There are roughly three types of approaches to machine translation:

1. **Rule-based methods** perform translation using extensive lexicons with morphological, syntactic and semantic information, and large sets of manually compiled rules. These systems are very labor intensive to develop.
2. **Statistical methods** entail the collection and statistical analysis of bilingual text corpora, i.e. parallel corpora. The technique tries to find the highest probability translation of a sentence or phrase among the exponential number of choices.
3. **Example-based methods** are similar to statistical methods in that they are parallel corpus driven. An Example-Based Machine Translator (EBMT) scans for patterns in both languages and relates them in a translation memory.

Most MT systems currently under development are based on methods (2) and/or (3). Research in these fields has greatly benefited from the increasing availability of parallel corpora, which are needed to bootstrap these approaches. Such a parallel corpus is typically aligned by linking, either automatically or manually, linguistic tokens in the source language to the corresponding units in the target language. Processing this data enables the development of fast and effective MT systems in both directions with a minimum of human involvement. In the next section we describe data collection and preprocessing efforts on the Sawa Corpus, a parallel corpus English – Kiswahili.

3. Data Collection And Annotation

While digital data is increasingly becoming available for Kiswahili on the Internet, sourcing useful bilingual data is far from trivial. At this stage in the development of the MT system, it is paramount to use faithfully translated material, as this benefits further automated processing. The corpus-based MT approaches we wish to employ, require word alignment to be performed on the texts, during which the words in the source language are linked to the corresponding words in the target language (also see Figure 1). But before we can do this, we need to perform sentence-alignment, during which we establish an unambiguous mapping between the sentences in the source text and the sentences in the target text. While some data is inherently sentence-aligned, other texts require significant preprocessing before word alignment can be performed.

The Sawa Corpus currently consists of a reasonable amount of data (roughly half a million words in each language), although this is not comparable to the resources available to Indo-European language pairs, such as the Hansard corpus [Roukos et al. 1997] (2.87 million sentence pairs). Table 1 gives an overview of the data available in the Sawa Corpus. For each segment it lists the number of sentences and words.

Table 1: Overview of the Data in the Sawa Corpus

	English Sentences	Kiswahili Sentences	English Words	Kiswahili Words
New Testament	7.9k		189.2k	151.1k
Quran	6.2k		165.5k	124.3k
Declaration of HR	0.2k		1.8k	1.8k
Kamusi.org	5.6k		35.5k	26.7k
Movie Subtitles	9.0k		72.2k	58.4k
Investment Reports	3.2k	3.1k	52.9k	54.9k
Local Translator	1.5k	1.6k	25.0k	25.7k
Full Corpus Total	33.6k	33.6k	542.1k	442.9k

We found digitally available Kiswahili versions of the New Testament and the uran for which we sourced the English counterparts. While religious material has a specific register and may not constitute ideal training material for an open-ended MT system, it does have the advantage of being inherently aligned on the verse level, facilitating further sentence alignment. Another typical bilingual text is the UN Declaration of Human Rights, which is available in many of the world's languages, including Kiswahili. This text was manually sentence-aligned.

The downloadable version of the on-line dictionary English-Kiswahili [Benjamin 2008] contains individual example sentences associated with the dictionary entries. These can be extracted and used as parallel data in the Sawa corpus. Since at a later point, we also wish to study the specific linguistic aspects of spoken language, we opted to have some movie subtitles manually translated. These can be extracted from DVDs and while the language is compressed to fit on screen and constitutes scripted language, they nevertheless provide a good sample of spoken language. It is inherently sentence-aligned, thanks to the technical time-coding information and also opens up possibilities for MT systems with other language pairs, since a commercial DVD typically contains subtitles for a large number of other languages as well.

The rest of the material consists of paragraph-aligned data, which was manually sentence-aligned. We obtained a substantial amount of data from a local Kenyan translator. Finally, we also included Kenyan investment reports. These are yearly reports from local companies and are presented in both English and Kiswahili. A major difficulty was extracting the data from these documents. The company reports are presented in colorful brochures in PDF format, meaning automatic text exports require manual post-processing and paragraph alignment. They nevertheless provide a valuable resource, since they come from a fairly specific domain and are a good sample of the type of text the projected MT system may need to process in a practical setting.

The reader may note that there is a very diverse range of texts within the Sawa corpus, ranging from movie subtitles to religious texts. While it certainly benefits the evaluation to use data from texts in one specific language register, we have chosen to maintain variety in the language data at this point. Upon evaluating the decoder at a later stage, we will however investigate the bias introduced by the specific language registers in the corpus.

All of the data in the corpus was subsequently tokenized, which involves automatically cleaning up the texts, conversion to UTF-8 and splitting punctuation from word forms. The next step involved scanning for sentence boundaries in the paragraph-aligned text, to facilitate the automatic sentence alignment method described in Section 4.

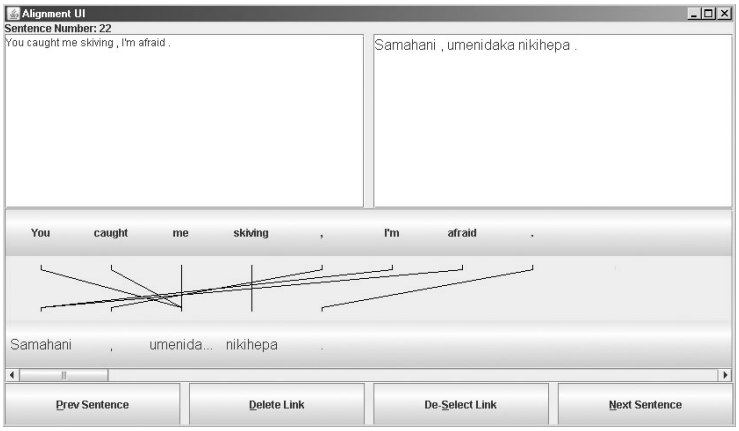
While not necessary for further processing, we also performed manual word-alignment annotation. This task can be done automatically, but it is useful to have a gold-standard reference against which we can evaluate the automated method. Monitoring the accuracy of the automatic word-alignment method against the

human reference, will allow us to tweak parameters to arrive at the optimal settings for this language pair.

We used the UMIACS word alignment interface [Hwa and Madnani 2004] for this purpose and asked the annotators to link the words between the two sentences (Figure 1). Given the linguistic difference between English and Kiswahili, this is by no means a trivial task. Particularly the morphological richness of Kiswahili means that there is a lot of convergence from words in English to words in Kiswahili. This alignment was done on some of the manual translations of movie subtitles, giving us a gold-standard word-alignment reference of about 5,000 words. Each annotator's work was cross-checked by another annotator to improve correctness and consistency.

4. Proposed Methods

Fig. 1: Manual word alignment using the UMIACS interface



There are a number of packages available to process parallel corpora. To preprocess the paragraph-aligned texts, we used Microsoft's bilingual sentence aligner [Moore 2002]. The output of the sentence alignment was consequently manually corrected. We found that 95% of the sentences were correctly aligned with most errors being made on sentences that were not present in English, i.e. instances where the translator decided to add an extra clarifying sentence to the direct translation from English. This also explains why there are more Kiswahili words in the paragraph aligned texts than in English, while the situation is reversed for the sentence aligned data.

For word-alignment, the state-of-the-art method is GIZA++ [Och and Ney 2003], which implements the word alignment methods IBM1 to IBM5 and HMM. While this method has a strong Indo-European bias, it is nevertheless interesting to see how far we can get with the default approach used in statistical MT. We evaluate by looking at the word alignments proposed by GIZA++ and compare them to the manually word-aligned section of the Sawa Corpus. We can quantify the evaluation by calculating precision and recall and their harmonic mean, the F-

score (Table 2). The former expresses how many links are correct, divided by the total number of links suggested by GIZA++. The latter is calculated by dividing the number of correct links, by the total number of links in the manual annotation. While the results presented in Table 2 are encouraging, it is clear that extra linguistic data sources and a more elaborate exploration of the experimental parameters of GIZA++ is needed.

Table 2: Precision, Recall and F-score for the word-alignment task using GIZA++

Precision	Recall	$F_{(ß=1)}$
39.4%	44.5%	41.79%

5. Discussion

In this paper we presented parallel corpus collection work that will enable the construction of a machine translation system for the language pair English – Kiswahili. We are confident that we have a critical amount of data that will enable good word alignment that can subsequently be used as a model for an MT decoding system, such as the Moses package [Koehn et al. 2007]. While the currently reported scores are not yet state-of-the-art, we are confident that further experimentation and the addition of more bilingual data as well as the introduction of extra linguistic features will raise the accuracy level of the proposed MT system.

The most straightforward addition is the introduction of part-of-speech tags as an extra layer of linguistic description, which can be used in word alignment model IBM5. The current word alignment method tries to link word forms, but knowing that for instance a word in the source language is a noun, will facilitate linking it to a corresponding noun in the target language, rather than considering a verb as a possible match. Both for English [Ratnaparkhi 1996] and Kiswahili [De Pauw et al. 2006], we have highly accurate part-of-speech taggers available. Another extra information source that we have so far ignored is a digital dictionary as a seed for the word alignment. The kamusiproject.org electronic dictionary will be included in further word-alignment experiments and will undoubtedly improve the quality of the output.

Once we have a stable word alignment module, we will further conduct learning curve experiments, in which we train the system with gradually increasing amounts of data. This will provide us with information on how much more data we need to achieve state-of-the-art performance. This additional data can be automatically found by parallel web mining, for which a few systems have recently become available [Resnik and Smith 2003]. Furthermore, we will also look into the use of comparable corpora, i.e. bilingual texts that are not straight translations, but deal with the same subject matter. These have been found to work as additional material within a parallel corpus [McEnery and Xiao 2007] and may further help improve the development of a robust, open-ended and bidirectional machine translation system for the language pair English - Kiswahili.

Acknowledgments

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). We are greatly indebted to Dr. James Omboga Zaja for contributing some of his translated data, to Mahmoud Shokrollahi-Far for his advice on the Quran and to Anne Kimani, Chris Wangai Njoka and Naomi Maajabu for their annotation efforts.

References

BENJAMIN, M. 2008. *The Kamusi Project.* Available at: http://www.kamusiproject.org (Accessed: 2 June 2008).

DE PAUW, G., DE SCHRYVER, G.-M. AND WAGACHA, P.W. 2006. Data-driven part-of-speech tagging of Kiswahili. In *Proceedings of Text, Speech and Dialogue, 9th International Conference* Springer Verlag, Berlin, Germany, 197-204.

DE PAUW, G. AND WAGACHA, P.W. 2007. Bootstrapping morphological analysis of Gikuyu using unsupervised maximum entropy learning. In *Proceedings of the eighth INTERSPEECH conference*, Antwerp, Belgium.

DE PAUW, G., WAGACHA, P.W. AND ABADE, D.A. 2007a. Unsupervised Induction of Dholuo Word Classes using Maximum Entropy Learning. In *Proceedings of the 1st International Conference in Computer Science and ICT (COSCIT 2007)* University of Nairobi, Nairobi, Kenya.

DE PAUW, G., WAGACHA, P.W. AND DE SCHRYVER, G.-M. 2007b. Automatic diacritic restoration for resource-scarce languages. In *Proceedings of Text, Speech and Dialogue, 10th International Conference* Springer Verlag, Berlin, Germany, 170-179.

HUTCHINS, W.J. 1986. *Machine translation: past, present, future*. Ellis, Chichester.

HWA, R. AND MADNANI, N. 2004. The UMIACS Word Alignment Interface. Available at: http://www.umiacs.umd.edu/ ~ nmadnani/alignment/forclip.htm (Accessed: 2 June 2008).

KOEHN, P., 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. AND 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.

MCENERY, A.M. AND XIAO, R.Z. 2007. Parallel and comparable corpora: What are they up to? In *Incorporating Corpora: Translation and the Linguist. Translating Europe. Multilingual Matters*, Clevedon, UK.

MOORE, R.C. 2002. Fast and Accurate Sentence Alignment of Bilingual Corpora. In *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users* Springer-Verlag, London, UK, 135-144.

OCH, F.J. AND NEY, H. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics 29*, 19-51.

RATNAPARKHI, A. 1996. A Maximum Entropy Model for Part-of-Speech Tagging. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, E. BRILL AND K. CHURCH Eds. Association for Computational Linguistics, Somerset, New Jersey, 133-142.

RESNIK, P. AND SMITH, N.A. 2003. The Web as a parallel corpus. *Computational Linguistics 29*, 349-380.

ROUKOS, S., GRAFF, D. AND MELAMED, D. 1997. *Hansard French/English.* Available at: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC95T20 (Accessed: 2 June 2008).