Xhosa-English Machine Translation: Working with a Low-Resource Language

Kristine K. Johnson
kkjohnson@wesleyan.edu

Summer 2011

Abstract

This report details the author's experiences as a Distributed Research Experience for Undergraduates (DREU) summer research intern at Carnegie Mellon University's Language Technologies Institute. Under the guidance of Prof. Carolyn Rosé, the author attempted to implement a phrase-based translation (i.e., statistical machine translation, or SMT) system for translating Xhosa text into English using the MOSES toolkit. Xhosa (*isiXhosa*) is a Bantu language widely used in South Africa; it is a *low-resource language* for which there is not a bevy of widely available language resources such as dictionaries, morphological analyzers, and so forth. Consequently, the author's work focuses heavily on the process of putting together a *parallel text*, or *bitext*, on which a working translation model could be trained.

1 Introduction

Translation is hard work: for the written word, there's the physical process of reading in text, mentally parsing it into comprehensible chunks, figuring out how words and phrases relate to one another or other larger linguistic structures, determining meaning from context... and then figuring out how to capture that meaning in another language, another way of thinking! Speech translation entails not only the aforementioned tasks, but also the (also difficult) task of signal processing to pick out words from what is essentially a stream of noise, but this is beyond the scope of our work. Given the

complexity of the translation problem, it is no wonder that computer-aided translation, which eases some of the burden, has become invaluable to human translators.[8] But what about fully-automated computer translation?

1.1 Machine Translation

The idea that a computer could translate between languages as well as a human might sound like something from Star Trek, and while results are far from perfect, the field of *machine translation* (MT) has made great leaps and bounds. In fact, MT has been around since around the birth of Artificial Intelligence (AI); in the 1970s, approaches such as direct modeling of language rules (i.e., *rule-based MT*) were attempted but proved to be rather unfruitful in terms of both results and the amount of time/linguistic expertise needed [11, 8]. Then in the 1990s, Peter F. Brown's group at IBM's Watson Laboratories published their seminal paper and introduced a promising mathematical formalization of the translation problem [3].

Brown, et al. introduced the paradigm of *statistical machine translation* (SMT), whereby a statistical translation model can be obtained from optimizing parameters from training data–*word-by-word alignments* from a *parallel text* consisting of sentences that are translations of one another in both the *target* and *source languages*–and generating a translation (some text in our target language, L2) of an input (some sentence in the source language, L1) by finding the most likely word alignments between L1 and L2 [3]. Acknowledging that "because of the minimal linguistic content of our algorithms, that it is reasonable to argue that word-by-word alignments are inherent in any sufficiently large bilingual corpus", MT research has largely focused around building and exploiting well such "sufficiently large" multilingual corpora, as well as developing algorithms to better estimate translation model parameters.

This is not to say that these are the only areas of interest in MT; indeed, recent efforts have brought MT full-circle, reinforcing the importance of linguistic knowledge and introducing language-based features such as synchronous context free grammars, syntax-augmented models, etc. [13] Moreover, there need not be a focus on only word-based alignments to produce good phrase tables. In fact, what is often meant by "SMT" is actually *phrase-based statis-*

tical machine translation, a generalization of word-based MT, wherein we are concerned with how not only single words, but strings of consecutive words (often up to 3 words, or *trigrams*) may correlate across the source and target languages [7].

1.2 Xhosa: A Low-Resource Language

The Xhosa language, or *isiXhosa*, is used predominantly in the Eastern Cape province of South Africa, and for numerous historical, social, and political reasons,[12] does not enjoy its due popularity in natural language processing (NLP) research; consequently, this presents the would-be NLP researcher with challenges unique to working with a *low-resource language*.

First and foremost, there is no well-researched, carefully compiled bilingual corpus of parallel Xhosa and English text (bitext) from which one may train a translation system. This presented the largest challenge to the author, who originally pursued the idea of using the bible as a training text–verses are already translated and aligned (give or take a few additions/omissions depending on the translation and edition), making the religious text a seemingly ideal candidate for use as a parallel training text (specifically, the open-source World English Bible, or WEB, and the 1975 Xhosa translation, both easily obtainable as plaintext online). However, compared to [9], the author lacked linguistic expertise in the source language (Xhosa), and also did not have a machine-readable bilingual (Xhosa-to-English) dictionary with which to supplement the biblical bitext. Moreover, the author consulted more than a few graduate students with considerable MT experience who strongly advised against using the bible as a training text, given its unique literary style and otherwise lack of resemblance to ordinary, everyday language (even with a more contemporary translation such as the WEB). Consequently, much work hard to be done to compile an appropriate corpus.

Additionally, Xhosa is a morphologically rich, highly agglutinative language, meaning that there is no shortage of prefixes, suffixes, and other affixes that may be used to modify a word, or stem [12]. For such languages, when attempting to translate a word–that despite existing in some form or another in the training text (and hence, phrase table)–chances are that it will appear to be *out-of-vocabulary* (OOV), and therefore be "untranslatable". Without

the benefit of morphological analysis, the only way to overcome this problem is by hopefully having a large enough corpus such that words in their many forms will appear. As a result, *data sparsity* remains a huge roadblock to any endeavor in SMT.

2 Methods

Given the difficulties in using the Bible as a parallel text for our purposes, the main challenge lied in finding a suitable resource from which the researcher could create a decent training text. Fortunately, a labmate pointed her towards a government website for the Western Cape province of South Africa, capegateway.gov.za (hereon referred to as "capegateway"), which provides general information in Xhosa, English, and Afrikaans, to the local populace about public health issues, municipal government structure, and other similar topics–nothing too *domain-specific*–providing the exact kind of "general", every-day use of language that is often desired in MT applications, particularly for the researcher's project (although arguably, when one knows the exact field or discipline in which the MT system is to be used, it is very much best to use an appropriate training text to capture the appropriate language use in language/translation models).

2.1 Extracting Parallel Text

Fortunately, capegateway's Xhosa and English websites exhibit parallel structure–all documents are named and organized the same way, just starting from different roots (`xh/` and `en/`, respectively). This made automatic extraction of (somewhat) aligned parallel text a very simple task. These webpages were obtained with wget using the mirror option (–M) to ensure that file/directory structure was preserved and we'd get as much data as possible, since in MT, it is often the case that "more data is *more*". After about a week of patient downloading, a large amount of data was collected, as can be seen in the following table. Note that duplicate files were removed.

Language	# of Files	Total Size
English	123,321	1.7 GB
Xhosa	133,902	1.9 GB
Afrikaans	133,851	1.8 GB

2.2 Cleaning Data

Unfortunately, capegateway's html took a great deal of effort to clean up. The author used `gawk` to eliminate html tags, which took not an insignificant amount of trial and effort to deal with the idiosyncracies of capegateway's html. After processing html pages with this filtering script, it was then necessary to use the `recode` tool to convert the text to UTF-8 encoding so that html escape sequences (mainly for the diacriticals in various government officials' names) would be properly processed. "Cleaner" text was collated into two files, `xho.txt` and `eng.txt`.

2.3 (Re-)Alignment

Unfortunately, there was a discrepancy of about 2000 lines of text between `xho.txt` and `eng.txt`, which resulted in MOSES aborting when attempting to train a baseline model. As a result, it was necessary to realign the text as best as possible using `hunalign`. However, since `hunalign` is not designed to handle extremely large datasets, it was necessary to chunk the text files to enable batch processing; again, unfortunately, even these chunked files were still too large to process, creating memory overflows, and it was necessary to rechunk those.

2.4 Other Pre-Processing

Boilerplate text (mostly from capegateway's navigation) were removed using `paste` and `cut`. The researcher also eliminated lines of text that contained no actual text; e.g., tables from fiscal reports, etc. Additionally, `hunalign` produces alignment scores, so two versions of the clean corpus were created: `filter-0` used 0 as a threshold value for alignment scores, thereby discarding blatantly bad alignments, and `filter-1`, which used 1.0 as a threshold to allow for more discriminative filtering. In the end, some 600,000 lines of parallel text were pared down to just a little over 4,000 lines. Finally, files that were not mirrored on both ends were recorded in a log file to hopefully be used as tuning and testing data. Thankfully, the researcher was able to (finally!) successfully train two corresponding baseline translation systems on these two versions of the capegateway corpus.

3 Results

A fellow DREU intern kindly provided the researcher with a csv file containing all instances of codeswitching from the Xhosa codeswitching corpus. Over 2000 lines of text representing instances of codeswitching (from both the General Conversation and School Talk subcorpora)[4] were used as input to both versions of the translation system. Not a single translation was found, although, being codeswitching instances, the system had to deal with a mix of English and Xhosa text.

4 Concluding Remarks

Despite producing less-than-satisfactory results, the overall project was a success in terms of introducing the researcher to the overall MT pipeline and useful tools. Furthermore, a decent–however small–Xhosa-English parallel corpus was produced. Had the researcher been able to successfully utilize her labmate's morphological analyzer, results would have undoubtedly improved. The author highly recommends utilizing morphological information in any future attempts to create a Xhosa-English translation system–and it also is incredibly helpful to have appreciable knowledge of South Africa and its languages, as it is possible to bootstrap working systems by exploiting similarities between languages such as Xhosa and Zulu.[2, 10, 1] In addition, performance may be improved by using language recognition to further clean a training corpus in pre-processing, particularly since not all of capegateway's supposedly Xhosa webpages were actually translated into Xhosa.

5 Acknowledgements

The author would like to express her gratitude toward the numerous individuals whose time and experience helped to make writing this paper possible: Jonathan Clark and his fellow PhD students in the LTI for sharing their invaluable expertise, David Adamson for his help with scraping data, Kaitlyn Price for her morphological analyzer and providing me with the pointer to the CapeGateway website (!!!), my fellow DREU intern Laura Willson for codeswitching csv file and her moral support, and most of all, Carolyn Rosé, for making the most wonderful and interesting summer of my life possible in the first place! Many thanks also go to the DREU program and its

hard-working administrators and mentors for reaching out to young female undergraduates in Computer Science.

Of course, the author would also like to recognize her family for their love and incomparable patience, Philippe and the Morels for their warm hospitality, and Londen and James for helping me get settled in Pittsburgh on the right foot.

References

[1] Allwood, J., Hammarström, H., Hendrikse, A., Ngcobo, M. N., Nomdebevana, N., Pretorius, L., and van der Merwe, M. Work on spoken (multimodal) language corpora in south africa. In *Proceedings of the International Conference on Language Resources and Evaluation* (2010), European Language Resources Association.

[2] Bosch, S. E., Pretorius, L., Podile, K., and Fleisch, A. Experimental fast-tracking of morphological analysers for nguni languages. In *Proceedings of the International Conference on Language Resources and Evaluation* (2008), European Language Resources Association.

[3] Brown, P. F., Pietra, S. D., Pietra, V. J. D., and Mercer, R. L. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics 19*, 2 (1993), 263–311.

[4] de Klerk, V. A. Codeswitching, borrowing and mixing in a corpus of xhosa english. *The International Journal of Bilingual Education and Bilingualism 9*, 5 (2006), 597–614.

[5] Koehn, P. *MOSES Statistical Machine Translation System User Manual and Code Guide*, 2011.

[6] 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. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics* (2007), Association for Computational Linguistics.

[7] Koehn, P., Och, F. J., and Marcu, D. Statistical phrase-based translation. In *Proceedings of the Human Languages Technology-North American Chapter of the Association for Computational Linguistics Conference (HLT-NAACL)* (2003).

[8] Manning, C. D., and Schütze, H. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[9] Phillips, J. D. The bible as a basis for machine translation. In *Proceedings of the Conference of the Pacific Association for Computational Linguistics* (2001), pp. 221–228.

[10] Pretorius, L., and Bosch, S. Exploiting cross-linguistic similarities in zulu and xhosa computational morphology. In *Proceedings of the First Workshop on Language Technologies for African Languages* (Stroudsburg, PA, USA, 2009), AfLaT '09, Association for Computational Linguistics, pp. 96–103.

[11] Russell, S., and Norvig, P. *Artificial Intelligence: A Modern Approach*, third ed. Prentice Hall, 2009.

[12] Webb, V., and Kembo-Sure, Eds. *African Voices: An Introduction to the Languages and Linguistics of Africa*. Oxford University Press, 2000.

[13] Zollmann, A., Venugopal, A., Och, F. J., and Ponte, J. M. A systematic comparison of phrase-based, hierarchical and syntax-augmented statistical mt. In *Proceedings of the 22nd International Conference on Computational Linguistics* (2008), pp. 1145–1152.