Experiments with syllable-based English-Zulu alignment

Gideon Kotzé, Friedel Wolff

University of South Africa
kotzegj@unisa.ac.za, wolfff@unisa.ac.za

Abstract

As a morphologically complex language, Zulu has notable challenges aligning with English. One of the biggest concerns for statistical machine translation is the fact that the morphological complexity leads to a large number of words for which there exist very few examples in a corpus. To address the problem, we set about establishing an experimental baseline for lexical alignment by naively dividing the Zulu text into syllables, resembling its morphemes. A small quantitative as well as a more thorough qualitative evaluation suggests that our approach has merit, although certain issues remain. Although we have not yet determined the effect of this approach on machine translation, our first experiments suggest that an aligned parallel corpus with reasonable alignment accuracy can be created for a language pair, one of which is under-resourced, in as little as a few days. Furthermore, since very little language-specific knowledge was required for this task, our approach can almost certainly be applied to other language pairs and perhaps for other tasks as well.
Keywords: machine translation, morphology, alignment

1. Introduction

Zulu is an agglutinative language in the Bantu language family. It is written in a conjunctive way which results in words that can contain several morphemes. Verbs are especially prone to complex surface forms. Although word alignment algorithms might have enough information to align all the words in an English text to their Zulu counterparts, the resulting alignment is not very useful for tasks such as machine translation because of the sparseness of morphologically complex words, even in very large texts. This is compounded by the fact that Zulu is a resource-scarce language.

A possible solution for this problem is to morphologically analyze each word and using the resulting analysis to split it into its constituent morphemes. This enables a more fine-grained alignment with better constituent convergence. Since verb prefixes often denote concepts such as subject, object, tense and negation, it would be ideal if they would align with their (lexical) counterparts in English. Figure 1 shows an example of a Zulu-English alignment before and after the segmentation. Here, it is clear that not only more alignments can be made, but in some cases, such as with *of*, we have better convergence as well.

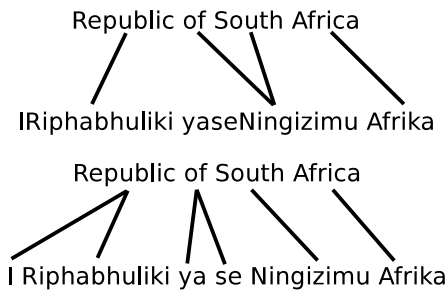

Figure 1: An example of an English-Zulu alignment before and after morphological segmentation.

Variations of this strategy have been followed with some success with similar language pairs such as English-Swahili (De Pauw et al., 2011) and English-Turkish (Çakmak et al., 2012).[1]

Morphological analysers are, however, difficult and time consuming to develop, and often relatively language specific. Although the bootstrapping of morphological analysers between related languages shows promise (Pretorius and Bosch, 2009), in each case the construction of a language-specific lexicon is still required, which is a large amount of work. The Bantu language family is considered resource-scarce, and methods that rely on technologies such as morphological analyzers, will mostly be out of reach for languages in this family.

We approach this problem by noting the fact that most languages in the Bantu family have a preference for open syllables (Spinner, 2011) and that in our case, even a simple syllabification approach can roughly approximate morphological segmentation. Hyman (2003) states that the open syllable structure of Proto-Bantu is reinforced by the agglutinative morphology. It is therefore possible to decompose words accurately for many Bantu languages into syllables in a straightforward way. If syllabification is useful for the task of word alignment (or indeed, any other task), it could be applicable to a large number of under-resourced languages. Indeed, some success has been demonstrated by Kettunen (2010) for an information retrieval task in several European languages. As far as we are aware, a syllable-based approach to alignment has not yet been implemented for Bantu languages.[2]

Figure 2 displays the previous example pair but with the words split into syllables instead of morphemes. Note that long proper nouns cause oversegmentation in the syllabification in comparison to its corresponding morphological segmentation. Since we have found this approach to work relatively well, we have, for the time being, decided not to segment the English texts morphologically.

[1] Turkish is also agglutinative.

[2] Some disjunctively written languages such as Northern Sotho (Griesel et al., 2010) and Tswana (Wilken et al., 2012), where the written words resemble syllables, have been involved in machine translation projects.

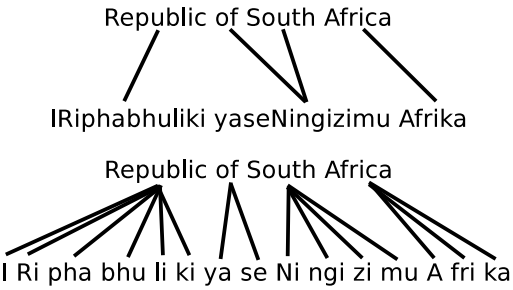

Figure 2: An example of an English-Zulu alignment before and after syllabification.

2. Data and preparation

For our experiments, we have attempted to obtain at least two different types of parallel text. Free-for-use Zulu-English texts are not so easy to find online, but eventually, we have chosen a marked-up version of the New Testament of the Bible (English: King James)[3] as well as the South African constitution of 1996.[4].

The Bible corpus is aligned on verse level. The fact that there are no abbreviations simplified the task of sentence splitting, which we deemed necessary since the length of verses may be too long for proper processing, especially in the case where words are split into morphemes. Therefore, we wrote and implemented a naive sentence splitter which assumes the lack of abbreviations.

Basic cleanup of the corpora was performed. In the Zulu Bible, we removed some extra text, as well as all double quotation marks, since they were not present in the English version. For English, we changed the first few verses to line up precisely with its Zulu counterpart to facilitate sentence alignment. In the constitution, we corrected some encoding issues. We also deleted some false translations and dealt with formatting-related issues such as tables which we removed.

Next, we used the above sentence splitter since the constitution also hardly contains any abbreviations. We then tokenized all the texts using the script *tokenizer.perl* which is distributed with Moses (Koehn et al., 2007), assuming no abbreviations.

Next, the sentence aligner Hunalign (Varga et al., 2005) was used to automatically align sentences. No dictionary was provided for the alignment process. In an attempt to ensure good quality output, only alignments with a probability score of 0.8 or above was used. We evaluated the alignment quality of a 5% sample (116 segments) of the aligned constitution and found only two problematic segments. In the first case, the sentence splitting was incorrect, whereas in the second case, a 10-token clause was omitted in the translation. We therefore feel confident that the alignments are of high quality.

While constructing a gold standard (see section 3.) we found that the alignment quality of the Bible corpus was poor. As such, we have decided not to use this for any quantitative evaluations. We suspect that differences in sentence composition, such as the handling of compound sentences (full stops or semi-colons in English versus commas in Zulu) have played a role.

Our last pre-processing step before invoking automatic word alignment was to segment the Zulu text into syllables. A very simple implementation was used where the end of the syllable is always assumed to be a vowel. This is a known rule in Zulu with few exceptions, such as in the case of loan words.[5]

Tables 1 and 2 show some statistics for each of the corpora.

3. Word alignment experiments and construction of gold standards

We invoked the unsupervised word aligner MGIZA++ (Gao and Vogel, 2008) on the sentence-aligned sentences. The output of both directions of alignment was combined with a selection of the heuristics as implemented in Moses (Koehn et al., 2007). Using this approach, we constructed a number of alignment sets, one for each method applied: *src2tgt* (source to target), *tgt2src* (target to source), *intersect* (intersection), *union*, *grow*, *grow-diag*, *grow-diag-final* and *grow-diag-final-and*. The set of *grow* heuristics are designed to balance precision and recall and work by iteratively adding links to the set of intersection alignments starting at neighbouring lexical units. For example, *grow-diag* focuses more on precision whereas *grow-diag-final* focuses more on recall. *src2tgt* refers to the asymmetrical source-to-target alignments of MGIZA++ where a source-side unit may only have one alignment but a target-side unit may have multiple, and with *tgt2src*, it is the other way around.[6]

Next, we proceeded to create small alignment gold standards for our corpora. Unfortunately, as mentioned before, the sentence alignment for the Bible corpus proved to be insufficient. Therefore, our gold standard only consisted of text from the constitution.

Our tool of choice was Handalign,[7] a tool for which, among other options, a graphical user interface can be used for the alignment of lexical units. We proceeded to correct output from the automatic alignments as combined with the *intersection* heuristic alignments, as this seemed like the method with the least amount of work. For this work, we did not make distinctions between high-confidence (*good*) and lower-confidence (*fuzzy*) alignments, although this would certainly be possible in the future.

As manual word alignment is non-trivial, we set about following a set of guidelines which we attempted to implement as consistently as possible. A few issues remain which we might address again in the future, depending on their influence on extrinsic evaluation tasks. For this experiment, we have decided to align as many units as possible for the facilitation of statistical machine translation. However, we still

[3] http://homepages.inf.ed.ac.uk/s0787820/bible/

[4] http://www.polity.org.za/polity/govdocs/constitution/

[5] One example of such an exception in the text is the Zulu word for *Sanskrit*, *isiSanskrit*, which was sillabified as *i si Sa nskri t*.

[6] We refer the reader to the following URL for more information: http://www.statmt.org/moses/?n=FactoredTraining.AlignWords

[7] http://www.cs.utah.edu/hal/HandAlign/

	Bible	Constitution	All
Sentence count	13154	3091	16245
Post-alignment sentence count	2245	2321	4566
Post-alignment/pre-syllabification token count	20628	33828	54456
Post-alignment/post-syllabification token count	58773	100619	159392

Table 1: Data statistics for the Zulu corpora. Sentence and token counts are valid for the texts after initial cleaning.

	Bible	Constitution	All
Sentence count	12535	3143	15678
Post-alignment sentence count	2245	2321	4566
Post-alignment token count	37244	45000	82244

Table 2: Data statistics for the English corpora. Sentence and token counts are valid for the texts after initial cleaning.

keep untranslated units with extra information, which may result in bad or unnecessary translations, unaligned. This includes function words and syllables. Additionally:

- In the case of non-literal translations, but for which clear boundaries still exist, such as between single words and phrases, the lexical units are still aligned. For example: *seat* (*of Parliament*) → *indawo yokuhlala* (literally: *place of sitting*)
- Where explicit counterparts for syntactic arguments exist, they are aligned. When, in Zulu, they are repeated in the form of syllabic morphemes, but no similar anaphor exists in English, we keep them unaligned. For example, in the case of *Money may be withdrawn* → *Imali ingakhishwa*, the first prefix *I-* is aligned with *Money* along with *-mali*. However, the subject concord in *ingakhishwa* (*i-*) which refers back to *Imali*, is not aligned. We have arrived at this decision based on the fact that such concords should align with English pronouns if present, but not with the antecedent noun. We thought that it would seem inconsistent if we decided to align the concord with the English noun only if the pronoun is not present. In the light of this, we have made the decision to not align the Zulu concords with the English nouns at all.
- In the case of phrases for which the segmentation into syllables and words makes no semantic sense (i.e. is too fine-grained), we attempt a simple and arbitrary monotic alignment. For example, with *take into account* → *bhe ke le le* and *Cape Town* → *Ka pa*, the word *take* is aligned with *bhe*, although the word *bhekelele* derived from the verb *bheka*, and *Cape* is aligned with *Ka* and *Town* with *pa*, although clearly no such distinctions exist.
- Where an English noun phrase of the form adj+noun was translated into a possessive noun phrase in Zulu, the possessive particle was not aligned. For example: *Electoral Commission* → *IKhomishani yokhetho* (literally: commission of voting). Here the syllable in the position of the possessive particle *yo-* was not aligned, since the English was not worded as a possessive noun phrase.
- Where an English noun phrase was translated with a Zulu noun phrase containing a relative clause, the relative prefix and optional suffix (-yo) were only aligned if an obvious English counterpart existed, such as *that* in the following example: *following words* → *amagama alandelayo* (literally: *words that follow*). Here the prefix *a-* and the suffix *-yo* are left unaligned as the English did not contain a relative clause.

For this work, we produced a small gold standard consisting of 20 sentence pairs. As this is too small to provide really meaningful quantitative results, we focus on a qualitative evaluation as a stepping stone to future alignment approaches.

4. Evaluation

Although we did not perform a quantitative evaluation of the Bible corpus, it may be worth noting that manual inspection suggests that proper nouns are frequently aligned succesfully. Eventually, this may prove to be useful for tasks such as named entity recognition or the compilation of proper name lexica.

The *tgttosrc* (target-to-source) combination heuristic only models one-to-many alignments from an English word to (possibly) multiple syllables in Zulu. A particularly interesting example of a successful alignment (even though with slight differences to our guidelines) is presented in figure 3. In this case the syllables of the noun *isakhiwo* (English: institution) are correctly aligned to the English noun, but also the cross alignments to the subject reference *si-* in *yisiphi* as well as the object reference *-si-* in *asinekeze* is correctly aligned.

The *intersection* heuristic provided the highest precision and lowest recall as expected. An interesting outcome from these alignments is that these alignments often selected the syllable in a Zulu noun or verb from the stem of the word. It therefore seems that this conservative heuristic is able to very accurately identify some kind of semantic "kernel" of the word:

- other → *enye* (aligned with *nye*)
- written → *esibhaliwe* (aligned with *bha*)

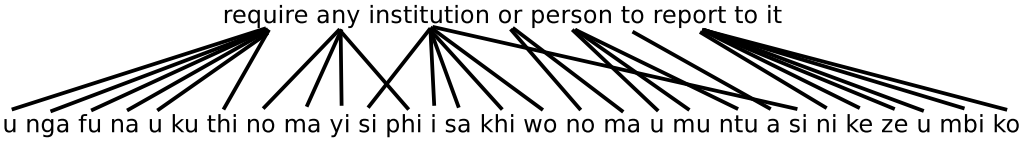

Figure 3: Example of automatic alignments generated by the *tgt2src* heuristic. This demonstrates the successful alignment of the noun *institution* with both the corresponding Zulu noun, as well as its corresponding subject and object *-si-* syllables. Both of these correspond to the proper morphological segmentation.

- person → *umuntu* (aligned with *ntu* — the monosyllabic noun stem)

The *union* alignment had the highest recall as expected. It also contained several incorrect long distance alignments and cross alignments.

Finally, for the sake of interest, we also provide precision, recall and F-score for the automatic word alignments as measured against the gold standard (Table 3).

The scores in relation to each are more or less expected. For example, *intersect* has the highest precision, *union* has the highest recall, while *grow-diag* has a higher precision but lower recall than *grow-diag-final*. However, the substantially higher recall of *tgt2src* in comparison with *src2tgt* is somewhat surprising. Although the high precision of *tgt2src* can be partly explained by the fact that the asymmetry of its alignment approximates our alignment approach, where a single English word is very often aligned to multiple Zulu syllables, it remains interesting that it almost has the highest F-score, beaten only by *grow-diag*. However, a larger gold standard is required to make any definite conclusions.

5. Future work

With the Zulu words now segmented into very fine constituent parts, the lack of similar segmentation in the English becomes more apparent. Although English is not agglutinative, some level of morphological analysis might still be useful. For example, past tense markers and plural suffixes are expected to align with certain syllables (or morphemes in the case of a morphological analysis). English prefixes such as *multi-*, *co-*, *re-*, *non-* are likely to find meaningful alignments with Zulu morphemes below the word level. Words with hyphens were not separated by the tokenizer. *Auditor-General*, *self-determination* and *full-time* are examples from the constitution corpus where simple splitting on hyphens could have made finer-grained alignment possible.

On the Zulu side, we would of course like to use an accurate morphological analyzer for the proper segmentation into morphological units. A promising candidate is ZulMorph (Pretorius and Bosch, 2003), which currently only outputs a list of candidate analyses.

In the long run, we hope to be able to create a larger gold standard comprising a variety of domains. With more training data, we should be able to train a decent machine translation system, although this certainly brings along its own set of challenges.

Another exciting prospect, especially considering the context of less-resourced languages, is the projection of English metadata such as POS tags and morpho-syntactic structure on the Zulu text in order to train taggers and parsers. For part-of-speech tagging, De Pauw et al. (2011) and Garrette et al. (2013) are among authors who have produced interesting work. For the projection of syntactic structure, see, for example, Colhon (2012).

6. Conclusion

Syllabification can be used successfully as a mostly language-independent method for word segmentation. For the task of word alignment, this facilitates more fine-grained word and morpheme alignment while not requiring the existence of a fully-trained morphological analyzer. Our work suggests that this can be applied succesfully to the English-Zulu language pair, requiring very little time and resources. We believe that this may provide opportunities for the faster development of resources and technologies for less-resourced languages, which includes the field of machine translation.

7. Data availability

For our experiments, we made use of data that are in the public domain. In the same spirit, we are making our processed data available under the Creative Commons Attribution-ShareAlike 4.0 licence (CC BY-SA 4.0). Please contact the authors for any inquiries.

8. References

Mehmet Talha Çakmak, Süleyman Acar, and Gülsen Eryigit. 2012. Word alignment for English-Turkish language pair. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2177–2180, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Mihaela Colhon. 2012. Language engineering for syntactic knowledge transfer. *Comput. Sci. Inf. Syst.*, 9(3):1231–1247.

Guy De Pauw, Peter Waiganjo Wagacha, and Gilles-Maurice Schryver. 2011. Exploring the SAWA corpus: collection and deployment of a parallel corpus English-Swahili. *Language Resources and Evaluation*, 45(3):331–344.

Qin Gao and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, SETQA-NLP '08, pages 49–57, Stroudsburg, PA, USA. Association for Computational Linguistics.

	Precision	Recall	F-score
intersect	**0.94**	0.28	0.43
grow	0.78	0.57	0.66
grow-diag	0.74	0.66	**0.699666**
grow-diag-final	0.62	0.75	0.68
grow-diag-final-and	0.72	0.68	0.697795
union	0.58	**0.76**	0.66
src2tgt	0.61	0.30	0.41
tgt2src	0.67	0.73	0.699216

Table 3: Precision, recall and F-score against the gold standard. Note how extremely close the top F-scores are to each other, and the interesting difference in recall between *src2tgt* and *tgt2src*.

Dan Garrette, Jason Mielens, and Jason Baldridge. 2013. Real-world semi-supervised learning of POS-taggers for low-resource languages. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL-2013)*, pages 583–592, Sofia, Bulgaria, August.

M. Griesel, C. McKellar, and D. Prinsloo. 2010. Syntactic reordering as pre-processing step in statistical machine translation from English to Sesotho sa Leboa and Afrikaans. In F. Nicolls, editor, *Proceedings of the 21st Annual Symposium of the Pattern Recognition Association of South Africa (PRASA)*, pages 205–210.

L.M. Hyman. 2003. Segmental phonology. In D. Nurse and G. Phillipson, editors, *The Bantu Languages*, pages 42–58. Routledge, New York.

Kimmo Kettunen, Paul McNamee, and Feza Baskaya. 2010. Using syllables as indexing terms in full-text information retrieval. In *Proceedings of the 2010 Conference on Human Language Technologies – The Baltic Perspective: Proceedings of the Fourth International Conference Baltic HLT 2010*, pages 225–232, Amsterdam, The Netherlands, The Netherlands. IOS Press.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Laurette Pretorius and Sonja E. Bosch. 2003. Finite-state computational morphology: An analyzer prototype for Zulu. *Machine Translation*, 18(3):195–216.

Laurette Pretorius and Sonja Bosch. 2009. Exploiting cross-linguistic similarities in Zulu and Xhosa computational morphology. In *Proceedings of the First Workshop on Language Technologies for African Languages*, AfLaT '09, pages 96–103, Stroudsburg, PA, USA. Association for Computational Linguistics.

Patti Spinner. 2011. Review article: Second language acquisition of Bantu languages: A (mostly) untapped research opportunity. *Second Language Research*, 27(3):418–430.

D. Varga, L. Németh, P. Halácsy, A. Kornai, V. Trón, and V. Nagy. 2005. Parallel corpora for medium density languages. In *Proceedings of Recent Advances in Natural Language Processing (RANLP-2005)*, pages 590–596.

I. Wilken, M. Griesel, and C. McKellar. 2012. Developing and improving a statistical machine translation system for English to Setswana: a linguistically-motivated approach. In A. De Waal, editor, *Proceedings of the 23rd Annual Symposium of the Pattern Recognition Association of South Africa (PRASA)*.