Morphological analysis for less-resourced languages: Maximum Affix Overlap applied to Zulu

Uwe Quasthoff[1], Sonja Bosch[2], Dirk Goldhahn[1]
[1]NLP Group, Dept. Comp. Sci., University Leipzig, Germany
[2]Dept. of African Languages, University of South Africa
E-mail: quasthoff@informatik.uni-leipzig.de, boschse@unisa.ac.za, dgoldhahn@informatik.uni-leipzig.de

Abstract

The paper describes a collaboration approach in progress for morphological analysis of less-resourced languages. The approach is based on firstly, a language-independent machine learning algorithm, Maximum Affix Overlap, that generates candidates for morphological decompositions from an initial set of language-specific training data; and secondly, language-dependent post-processing using language specific patterns. In this paper, the Maximum Affix Overlap algorithm is applied to Zulu, a morphologically complex Bantu language. It can be assumed that the algorithm will work for other Bantu languages and possibly other language families as well. With limited training data and a ranking adapted to the language family, the effort for manual verification can be strongly reduced. The machine generated list is manually verified by humans via a web frontend.

Keywords: Complex morphology, Zulu, machine learning algorithm

1. Introduction

The paper describes work in progress. A two-step process is used to generate high quality morphological data for less-resourced languages, especially in the case of languages with complex morphologies. In these cases one cannot expect to have an automatic high quality analysis without extensive training data. The training data are usually expensive and for many languages, cannot be generated.

The approach introduced here is explained for morphological decomposition, but is applicable for solving other challenges as well (as described in section 5). We proceed as follows:

(1) Startingfrom an initial set of training data (i.e. words with their morphological decomposition), a machine learning algorithm generates candidates for morphological decompositions of 'new' words. This training set may be relatively small and also may contain errors or other inconsistencies. For each word, this might be either one in a ranked list of possible decompositions or just one (i.e. the most probable) decomposition.

(2) The machine generated list is manually verified by humans via a web frontend. Their task is to mark the correct decompositions. Alternatively, a word can be marked as "incorrectly analyzed" and the correct analysis can be inserted. A typical result is one correct decomposition per word. In the case of ambiguities, several decompositions might be correct. A word is treated as verified if at least one decomposition is marked as correct or an additional decomposition has been added. It is treated as not verified (and will be presented to another person for verification later) if nothing is marked.

The quality of both the annotated data and machine generated decompositions can be increased using a more complicated process:

- For higher quality and/or measuring agreement of different annotators, some or all entries can be presented to several persons. Additionally, pattern based algorithms may search for inconsistencies in the annotated data.
- The results of the human verification can be regarded as additional training data, with the result that the quality of the data presented in (2) increases steadily.

It should be noted that the task described in (2) is much simpler than decomposition without any suggestions. Choosing from a set of alternatives is less time consuming and needs less proficiency. For these reasons the task is well suited for a collaboration scenario.

The procedure above is demonstrated on Zulu morphology which is representative of many languages with complex morphology: morphological analysis is a prerequisite for POS tagging due to numerous short affixes and roots of possibly only one character.

2. Complex morphology of Zulu

Zulu [ISO 639-3: zul] belongs to the family of Bantu languages which have a complex morphological structure, based on two principles: a nominal classification system, and a concordial agreement system. According to the nominal classification system, nouns are categorized by prefixal morphemes that have been given class numbers for analysis purposes. These noun class prefixes generate concordial agreement linking the noun to other words in the sentence such as verbs, adjectives, pronouns, possessives etc. (cf. Poulos and Msimang, 1998) as illustrated by the bold printed morphemes in the following sentence:

> ***Aba**ntu **aba**ningi **ba**nga**yi**chitha **i**mali **ya**bo.*
> **Aba**-ntu **aba**-ningi **ba**-nga-**yi**-chitha **i**-mali **ya**-bo.
> [Many people may waste their money.]

In this example, the class 2 noun *abantu* [people] determines the subject agreement morpheme*ba-* in the verb *bangayichitha* [they may waste it], as well as the adjective agreement *aba-* in the qualificative*abaningi* [who are many]. The class 9 noun *imali* [money] determines object agreement *-yi-* in the verb and possessive agreement *-ya-* in *yabo* [of them]. We follow the root-based approach in morphological analysis of

Zulu where the root carries the principal semantic load of the word, e.g. *-ntu* and *-mali* (noun roots) and *-chith-* (verb root) in the sentence above. It should be noted that noun and verb roots belong to an open class which may demonstrate continuous growth.

The conjunctive orthography of the Zulu language causes a certain degree of morphophonological complexity. Most of the phonological adjustments at morpheme boundaries are predictable and rule-based. However, there are some exceptions - these are handled in the training data.

Zulu as a less-resourced language

According to Scannell (2007:1), more than 98% of the world's living languages lack most of the basic resources needed as a base for advanced language technologies, and are referred to as less-resourced or under-resourced languages. Zulu can therefore also be regarded as a less-resourced language, considering the unavailability of e.g. large monolingual and bilingual corpora, machine-readable dictionaries, POS taggers, morphological analysers, parsers, etc. Although some corpora exist (cf. University of Pretoria[1], Language Resource Management Agency[2] and Leipzig Corpora Collection[3]), they are limited in size, are not annotated and often not even accessible. Morphological analysers for Zulu are reported on e.g. a finite-state morphological analyser ZulMorph (Bosch et al. 2008), machine learning Zulu analysers (Spiegler et al. 2008; Shalanova et al. 2009), and a bootstrapping approach (Joubert et al. 2005). However, none of these morphological analysers is freely available.

The following algorithm describes a morphological analyser with a strict separation of the language-independent algorithm and the language specific training data. It can be assumed that the algorithm will work for other Bantu languages and possibly other language families as well. Possible language dependent limitations will be treated in a post processing step.

In the following section the algorithm is described without special reference to Zulu. Only the examples are taken from this language.

3. Morphologic Decomposition: The Maximum Affix Overlap Algorithm

Algorithms for morphological decomposition can use training data (so-called supervised algorithms) or use only word forms without any additional information (unsupervised algorithms like Morfessor (Creutz et al. 2006)). Unsupervised algorithms often have problems with complex morphologies; therefore we chose a supervised algorithm. The repeated succession of some of the morphemes will be used to classify the morphemes using the training data. In contrast to a rule-based morphological analyser such as ZulMorph (Bosch et al. 2008) that depends on a word root lexicon for successful analyses, this approach concentrates on affixes and allows the identification of previously unknown roots. The only additional assumption is that there is exactly one central element in the word, namely the root. In the case of compounds with two or more roots we assume that compound decomposition was applied in advance. We do not assume that the morphological analysis is unique. Instead, both the segmentation and the classification of the segments may be ambiguous.

Step 1: Language independent decomposition and ranking

We start with training data containing decompositions for a certain number of words so that we can assume that all possible combinations of prefixes are contained in the data as well as all combinations of suffixes. We do not assume, however, that all combinations of prefixes and suffixes are contained in the training data. Moreover, we do not assume all roots to be known in advance because one of the aims is to detect unknown or 'new' roots. In fact, checking for known roots will be postponed for step 2 of the algorithm. The algorithm returns a ranking of different decompositions and tags which is appreciated for languages with complex morphology because multiple decompositions are possible.

For a word *w* to be analysed, we perform the following steps:

For all segmentations of the word *w* into three segments *w1*, *w2* and *w3* (where *w1* and *w3* might be of zero length) we do the following:

- For each word *x* in the training set having exactly the prefix sequence *w1* we collect the pair (morphological analysis of *w1*, *w2* with the tag of the root of *x*).
- For each word *x* in the training set having exactly the suffix sequence *w3* we collect the pair (*w2* with the tag of the root of *x*, morphological analysis of *w3*).

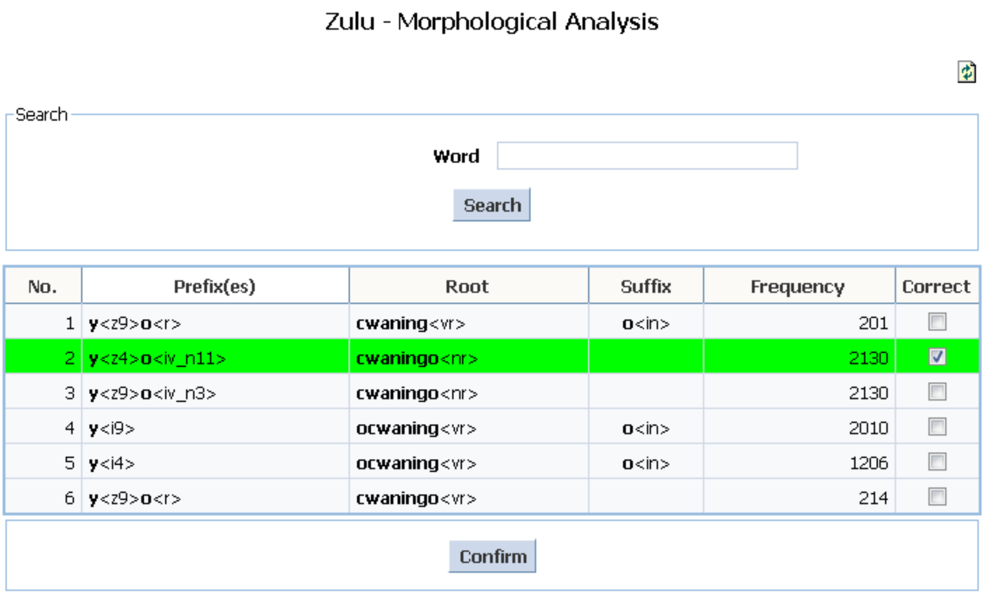
From this, we form triples by joining on identical root tags: (morphological analysis of *w1*, *w2* with the tag of the root of *x*, morphological analysis of *w3*). Interesting features are the length of *w2* and the frequency of identical triples above. Because the procedure above allows considering affixes next to the root as part of the root, shorter roots should be preferred. In the case of multiple decompositions with the same root (or different roots of the same length) we rank the decompositions according to the frequency of the corresponding triple (morphological analysis of *w1*, tag of the root of *x*, morphological analysis of *w3*). In general, we set a frequency threshold of 2 for decompositions to be considered.

The example in Table 1 shows the analysis of the word *yocwaningo* [of research]. The correct decomposition has the highest frequency, but a shorter root candidate ranks higher.

[1] http://web.up.ac.za/default.asp?ipkCategoryID=1866&subid=1866&ipklookid=9

[2] http://rma.nwu.ac.za/

[3] http://corpora.informatik.uni-leipzig.de/



No.	Prefix(es)	Root	Suffix	Frequency	Correct
1	y<z9>o<r>	cwaning<vr>	o<in>	201	☐
2	y<z4>o<iv_n11>	cwaningo<nr>		2130	☑
3	y<z9>o<iv_n3>	cwaningo<nr>		2130	☐
4	y<i9>	ocwaning<vr>	o<in>	2010	☐
5	y<i4>	ocwaning<vr>	o<in>	1206	☐
6	y<z9>o<r>	cwaningo<vr>		214	☐

Table 1: Analysis of the word *yocwaningo* in the verification tool: line no. 2 is correct.

Step 2: Language dependent post-processing using special patterns
Step 1 of the algorithm produces too short roots if possible affixes (or parts thereof) are instead part of the root. In this case, it is not the shortest root candidate that will be the correct one. Here some language specific patterns help to exclude root candidates or give them a lower rating:

- Roots might not begin or end with some character or character sequence.
- Some roots can be extended by one character (usually a vowel) which also might be a suffix.
- Blacklisting: Some incorrect very short root candidates will be generated repeatedly. They can be blocked using a blacklist.
- Root transformations: The algorithm fails if the root is not part of the input word. But for Zulu, this happens only in rare cases. The most frequent transformation rule is given here: In a case such as the locative noun *ezandleni* [in the hands] the algorithm incorrectly provides *-andl-* as noun root. The locative prefix *e-* is necessary toensure that *-eni* is indeed a locative suffix and, moreover, that the nounroot is ending in *-a* or *-e*. Hence, this rule generates two possible roots: *-andla* (correct) and *andle* (incorrect).
- Agreement: In some cases, agreement between the prefix tag and the root is required. The noun root in the word *nomndeni* [and the circle of relatives] seems on the surface, to have a locative suffix *-eni*, and therefore the correct noun root-*ndeni* [circle of relatives] (class 3 noun) is not recognised. However, the absence of a locative prefix *e-* is the clue to the fact that there cannot be a locative suffix in this noun. Hence, *-eni*is part of the root.

Using frequency data for re-ranking
The following rules can be used if we have frequency information for roots. Usually, higher frequency should give a higher ranking. Such frequency information is:

- Frequency of a root in the training data (always available)
- Frequency of a root candidate (usually if not in the training data) in analysed corpus data.If, for instance, a correct root might be extended with several different vowels, these extensions will automatically get lower frequencies.

Training data
The training data used is the Ukwabelana (2013) word list consisting of approx. 10,000 words with labelled analyses described in Spiegler et al. (2010).

Evaluation
For 50 words of medium frequency (of frequency class 7, i.e. the most frequent word *ukuthi* [that / so that] is about 2^7 as frequent as the test words), which were randomly selected from a Zulu Newspaper Corpus of the Leipzig Corpora Collection [4], the automatic analyses were manually checked for the first correct analysis. It is counted for how many words the first analysis is correct, a correct analysis is in the top-5 or top-10 analyses provided. Here, both correctness of full analysis and correctness only for the root and its type are distinguished (cf. Table 2).

	Complete analysis		**Only root and its type**	
	absolute	**%**	**absolute**	**%**
total	50	100%	50	100%
correct at pos. 1	26	52%	30	60%
correct within pos. 1-5	32	64%	36	72%
correct within pos. 1-10	41	82%	41	82%
not correct within pos. 1-10	9	18%	9	18%

Table 2: Evaluation of the Maximum Affix Overlap algorithm

[4] http://corpora.informatik.uni-leipzig.de/

4. Sample Application: Identifying new roots

Here we focus in particular on the open classes of morphemes, viz. nounand verb roots. The root guesser using the above methods will facilitate the identification of "new" or adopted noun and verb roots that do not as yet occur in existing dictionaries or lexicons of the language. Such lists of "new" roots can be shared and integrated into existing applications such as ZulMorph and at the same time contribute to language development and orthographic standardisation purposes.

Example: The noun root *cwaningo* [research] (noun class 11) is a derivation from the verb root -*cwaning*- [conduct research] that does not feature in most Zulu dictionaries since it is a relatively "new" coinage. The correct analysis can be found in the Table 1 above.

5. Future work

It is planned to test the Maximum Affix Overlap algorithm for the other official Bantu languages of South Africa. Training data should become available in 2014 from the Language Resource Management Agency[5]. For future work, a more elaborate tag set than that used in (Spiegler et al. 2008, 2010) will be considered since the output of the analysis should be suitable for a POS-Tagger.

Both software and data for additional languages will be made availableunder the creative commons license by-nc. The procedure described for morphological decomposition can also be applied to other tasks. The common feature is that for each input word the correct output has to be generated using machine learning and manual correction. This scenario applies to several problems such as the following:

- inflection type / baseform reduction, morphological decomposition, compound decomposition
- classification tasks for subject areas or relations (as in WordNet)
- bilingual translation equivalents

The combined data created for several of the above problems can contribute to improve the quality of the machine generated data.

6. References

Bosch, S., Pretorius, L. and Fleisch, A. (2008). Experimental Bootstrapping of Morphological Analysers for Nguni Languages.Nordic Journal of African Studies 17(2):66-88. Available: http://www.njas.helsinki.fi/. Accessed on 20 February 2014.

Creutz, M., Lagus, K. and Virpioja, S. (2006). Unsupervised morphology induction using Morfessor. In A. Yli-Jyrä, L. Karttunen, and J. Karhumäki (Eds.), Finite-State Methods and Natural Language Processing, Finite-State Methods and Natural Language Processing (FSMNLP 2005), Volume 4002 of Lecture Notes in Computer Science, pp. 300–301. Berlin, Heidelberg: Springer-Verlag.

Joubert, L., Zimu, V., Davel, M. and Barnard, E. (2004). A framework for bootstrapping morphological decomposition. Available: http://www.meraka.org.za/pubs/joubertl04morphanalysis.pdf. Accessed on 12 October 2013.

Poulos, G. and Msimang, C.T. (1998). A linguistic analysis of Zulu. Pretoria: Via Afrika.

Scannell, K.P. (2007). The Crúbadán Project: Corpus building for under-resourced languages. In C. Fairon, H. Naets, A. Kilgarriff, and G-M. de Schryver (Eds). Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop. Cahiers du Cental, Louvain-la-Neuve, Belgium, Vol. 4 pp. 5-15.

Shalonova, K., Golenia, B. and Flach, P. (2009). Towards learning morphology for under-resourced languages. IEEE Transactions on Audio, Speech and Language Processing, 17(5):956–965.

Spiegler, S. (2011). Machine Learning for the Analysis of Morphologically Complex Languages. PhD Thesis. Merchant Venturers School of Engineering, University of Bristol.

Spiegler, S., Golenia, B., Shalonova, K., Flach, P. and Tucker, R. (2008). Learning the morphology of Zuluwith different degrees of supervision. IEEE Spoken Language Technology Workshop, pp. 9–12.

Spiegler, S., van der Spuy, A. and Flach, P.A. (2010). Ukwabelana - An open-source morphological Zulu corpus. Proceedings of the 23rd International Conference on Computational Linguistics (COLING), pp. 1020–1028.

Ukwabelana - An open-source morphological Zulucorpus. (2013). Available: http://www.cs.bris.ac.uk/Research/MachineLearning/Morphology/resources.jsp. Accessed on: 11 October 2013.

[5] http://rma.nwu.ac.za/