Heuristic guided probabilistic graphic language modelling for morphological segmentation of isiXhosa

Lulamile Mzamo[1], Albert Helberg[1] and Sonja Bosch[2]

[1]North-West University and [2]UNISA, South Africa.

Abstract

The IsiXhosa Heuristics Maximum Likelihood Segmenter (XHMLS), an unsupervised isiXhosa segmenter, that contributes use of isiXhosa word morphology heuristics as a guide to probabilistic graphical modelling (PGM) in the segmentation of isiXhosa, outperforms the benchmark Morfessor-Baseline's boundary identification (BI) accuracy of 77.2 ± 0.10%, by 1.5 ± 0.01% and its BI f1-Score also outperforms Morfessor-Baseline's 48.9 ± 0.75% by 19.1 ± 0.06% when modelled with circumfixing and modified Kneser-Ney (mKN) smoothing.

I-IsiXhosa Heuristics Maximum Likelihood Segmenter (XHMLS), isicaluli-mbhalo sesiXhosa esingagadwanga, esigalelo ikukusetyenziswa kwendlela amagama esiXhosa aguquka ngayo njengesikhokelo somFanekiso-mBoniso-Thuba (FBT) ekucaluleni isiXhosa, igqitha inkcaneko yokukhomba imida (KD) yezimilo yomgangatho-jikelele oyiMorfessor-Baseline, eyi-77.2 ± 0.10%, nge-1.5 ± 0.01%, yaye inqaku le-f1 leKD layo liligqithe ngakumbi eleMorfessor-Baseline elingu-48.9 ± 0.75%, nge-19.1 ± 0.06%, xa inkokhelo izizimi-macala yaye igudiswa nge-Kneser-Ney elungisiweyo (mKN).

Introduction

This paper details and evaluates the IsiXhosa Heuristics Maximum Likelihood Segmenter (XHMLS) against the benchmark Morfessor-Baseline [1] in terms of morpheme boundary identification accuracy and f1-score.

XHMLS is an unsupervised heuristic guided probabilistic graphical modelling (PGM) [2] based segmenter that implements four (4) isiXhosa word generation models.

Morphological analysis, the task that XHMLS attempts to solve, is one of the basic tools in the natural language processing (NLP) of agglutinating languages, like isiXhosa.

IsiXhosa, one of the South African official languages belonging to the Bantu language family, is classified among the "resource scarce languages". The second largest language in South Africa with 8.1 million mother-tongue speakers (16% of the South African population), second only to isiZulu [3], isiXhosa has seen an increase in HLT tools, however this increase has been from a low baseline [4].

The close morphological structure that isiXhosa has with other Nguni languages, i.e. isiZulu, Siswati and isiNdebele, means that work done in it could be easily bootstrapped to these languages as has been shown in [5]. Nguni languages account for 43% of the South African mother tongue speaker population.

IsiXhosa Word Constitution

Morphological segmentation is the task of splitting a word, into its constituent smallest meaning bearing units, the morphemes [6].

IsiXhosa is an agglutinating and polysynthetic language as its words are made up of many morphemes [7], e.g.

akahambanga <- a-ka-hamb-ang-a (he/she did not go).

XHMLS IsiXhosa Morphology Heuristics

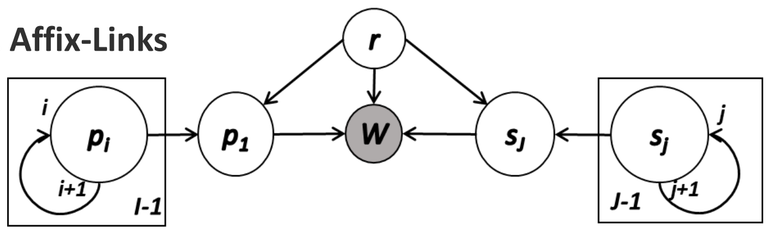
The following heuristics are used to constrain XHMLS's segmentation search space:

- If the first character is a vowel it is always a morpheme;
- If the last character is a vowel it is always a morpheme (terminal vowel);
- Prefix morphemes are complete isiXhosa syllables, except for *m* which has a silent vowel when followed by a consonant;
- Suffix morphemes start with a vowel and end in a consonant except for the terminal vowel and for *w* which has a silent preceding vowel when following a consonant;
- roots start with a consonant.

Lulamile Mzamo (Lula_Mzamo@yahoo.co.uk)
Albert Helberg (Albert.Helberg@nwu.ac.za)
Sonja Bosch (Boschse@unisa.ac.za)

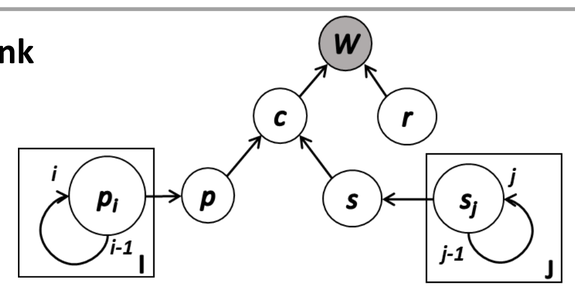
XHMLS Models

Fig 1. Affix-Links



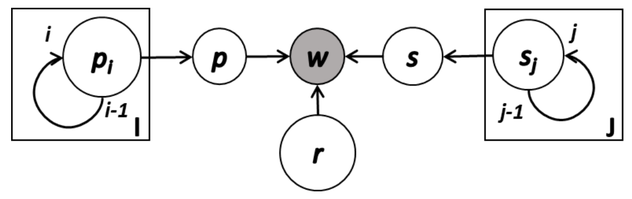
$$P(\boldsymbol{S}) = P(r) \cdot P(p_1|r, p_2) \cdot \prod_{i=2}^{I} P(p_i|p_{i-1:1}) \cdot P(s_J|r, s_{J-1}) \cdot \prod_{j=1}^{J-1} P(s_j|s_{j+1:J}) \quad (1)$$

Fig. 2 Circumfix-Link



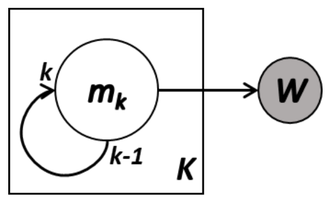
$$P(\boldsymbol{S}) = \frac{P(c)}{P(r) \cdot \prod_{i=1}^{I} P(p_i|p_{i-1:1}) \cdot \prod_{j=1}^{J} P(s_j|s_{j-1:1})} \quad (2)$$

Fig. 3 Independent Affixes



$$P(\boldsymbol{S}) = P(r) \cdot \prod_{i=1}^{I} P(p_i|p_{i-1:1}) \cdot \prod_{j=1}^{J} P(s_j|s_{j-1:1}) \quad (3)$$

Fig. 4 Morpheme Sequences



$$P(\boldsymbol{S}) = \prod_{k=1}^{K} P(m_k|m_{k-1:1}) \quad (4)$$

where:

- ***W*** is a word
- ***S*** is a word segmentation
- *r* is a word root
- *c* is a circumfix
- *p* is a prefix
- *s* is a suffix
- p_i is the *i*-th prefix morpheme, and
- s_j is the *j*-th prefix morpheme
- *I* is numer of prefix morphemes
- *J* is the number of suffix morphemes
- m_k is the *k*-th morpheme, and
- *K* is number of morphemes in a word
- *P(·)* is the probability of an event
- *P(·|<unk>) ≡ P(·)*

Results

Method	Accuracy (%)	F1-Score (%)
Affix Links	73.6 ± 0.05	59.2 ± 0.16
Circumfix Links	**78.7 ± 0.01**	**68.0 ± 0.06**
Independent Affixes	75.7 ± 0.03	58.3 ± 0.51
Morpheme Sequences	74.7 ± 0.24	59.5 ± 0.30
Morfessor-Baseline	77.2 ± 0.10	48.9 ± 0.75

Conclusions

The use of morphology heuristics in isiXhosa segmentation is feasible.

The effect of circumfixes in the modelling of isiXhosa improves the performance of a guided probabilistic graphical model, but that depends on the generative model used.

References

1. Creutz, M., Lagus, K.: Unsupervised Discovery of Morphemes. In: Procedings of the 6th Workshop of the ACL Special Interest Group in Computational Phonology. pp. 21–30. , Philadelphia, USA (2002).
2. C. M. Bishop, Pattern Recognition and Machine Learning. Springer, 2006.
3. Statistics South Africa.: Census 2011: Census in brief. (2012).
4. Moors, C., Calteaux, K., Wilken, I., Gumede, T.: Human language technology audit 2018: Analysing the development trends in resource availability in all South African languages. In: SAICSIT 2018. pp. 296–304. ACM, Port Elizabeth, South Africa (2018). https://doi.org/10.1145/3278681.3278716.
5. Bosch, S., Pretorius, L., Fleisch, A.: Experimental Bootstrapping of Morphological Analysers for Nguni Languages. Nord. J. African Stud. 17, 66–88 (2008).
6. Manning, C.D., Schütze, H.: Foundations of Statistical Natural Language Processing. MIT Press (1999). https://doi.org/10.1162/coli.2000.26.2.277.
7. Kosch, I.M.: Topics in Morphology in the African Language Context. Unisa Press, Pretoria (2006).

LT4ALL