Statistical morpheme segmentation in Inuktitut

Jessica Huynh

December 18, 2016

1 Introduction

Inuktitut is an Inuit language spoken mostly in Canada, and like many other Eskimo-Aleut languages, it is highly agglutinative, although not very fusional.[1]

2 Background

The highly regular nature of Inuktitut morphology means that one can write a rule-based morpheme segmenter with relatively high accuracy. Indeed, the National Research Council of Canada has already built one, claiming over 95% accuracy on the Nunavut Hansard corpus and similar accuracy on Inuktitut web pages, composing a list of thousands of roots and suffixes to do so.[3]

3 Experiment

Morfessor[4] is a package of

3.1 Method

[4]

3.2 Approaches

3.3 Corpora

For the corpora, I used the Nunavut Hansard corpus[2]

3.4 Tools

4 Results

5 Discussion

6 Conclusion

References

[1] Inuktitut @ omniglot. http://www.omniglot.com/writing/inuktitut.htm.

[2] National Research Council of Canada. The nunavut hansard: Inuktitut-english parallel corpus. http://www.inuktitutcomputing.ca/NunavutHansard/info.php?lang=en, 2008.

[3] National Research Council of Canada. The uqailaut project: Inuktitut morphological analyzer. http://www.inuktitutcomputing.ca/Uqailaut/info.php, 2012.

[4] Sami Virpioja, Peter Smit, Stig-Arne Grönroos, and Mikko Kurimo. *Morfessor 2.0: Python Implementation and Extensions for Morfessor Baseline.*