NLP Tools for Khasi, a low resource language

Medari Janai Tham
Department of Computer Science and Engineering
Assam Don Bosco University, Assam, India
medaritham16@gmail.com

Abstract

Khasi is an Austro Asiatic language spoken by one of the tribes in Meghalaya, and parts of Assam and Bangladesh. The fact that some NLP tools for Khasi are now available online for testing purposes is the culmination of the arduous investment in time and effort. Initially when work for Khasi was initiated, resources for Khasi, such as tagset and annotated corpus or any NLP tools, were nonexistent. As part of the author's ongoing work for her doctoral program, currently, the resources for Khasi that are in place are the BIS (Bureau of Indian Standards) tagset for Khasi, a 90k annotated corpus, and NLP tools such as POS (parts of speech) taggers and shallow parsers. These mentioned tools are highlighted in this demonstration paper.

1 Introduction

Khasi can be categorized as a low resource language from language technology standpoint due to the fact that reported resources available for Khasi such as a 90k annotated corpus is comparatively small in size (Tham 2020b). However, this did not hinder the development of POS (parts of speech) taggers and shallow parsers for Khasi where their performances are at par with performances of other reported taggers and parsers of other Indian languages. These tools are now available for testing online in https://medaritham.pythonanywhere.com, where users can enter a sentence in Khasi and observe and compare the tagging and parsing performances of various techniques employed in the tools.

2 Khasi BIS Tagset and Annotated Corpus

Due to the unavailability of any corpus in Khasi, the required corpus has to be built from scratch which consumes time and effort. The corpus size is 94,651 tokens extracted from across 38 stories of the prose genre. Details on Khasi corpus construction are explicated in (Tham 2018b) and (Tham 2020b). Next the Khasi tagset is formulated according to the BIS (Bureau of Indian Standards) tagset and the total number of tags is 33. Likewise, further details on Khasi tagset formulation can be found in (Tham 2018b). With the availability of these resources, POS taggers for Khasi were designed and a brief description is given in Section 3.

3 POS Taggers for Khasi

Initially, a Hidden Markov Model (HMM) POS tagger for Khasi was developed and tested only on test data comprising of text from a book not seen during training and achieved an accuracy of 95.68% (Tham 2018b). However rigorous testing of the HMM POS tagger was carried out using ten-fold cross validation giving an accuracy of 93.39%. In order to address the tagging errors of the HMM POS tagger, a Hybrid POS tagger for Khasi was developed which reported an accuracy of 95.29% using ten-fold cross validation (Tham 2020b). This was possible due to the integration of conditional random fields (CRF) which allows the incorporation of language features not possible in an HMM POS tagger. The language features included for Khasi are capitalization, prefixes, current word under consideration, previous word, next word, and whether a word begins or ends a sentence. Due to the absence of inflection, prefixes

Proceedings of the 17th International Conference on Natural Language Processing: System Demonstrations, pages 26–27
Patna, India, December 18 - 21, 2020. ©2019 NLP Association of India (NLPAI)

are prevalent in Khasi exhibiting derivational morphology. An additional feature included is the previous tag of a word. Both these POS taggers can be observed and compared online in the site mentioned earlier. One distinct difference is the ability of the Hybrid tagger to differentiate proper nouns from common nouns a trait where CRFs excel.

4 Shallow Parsers for Khasi

To train a shallow parser for Khasi, the annotated corpus has to be further tagged with noun and verb chunks using the BIO labelling specified by Ramshaw and Marcus (1995) where each alphabet symbolizes the following:

B-XX: label **B** for a word starting a chunk of type XX.
I-XX: label **I** for a word inside a chunk of type XX.
O: label **O** for a word outside of any chunk.

Shallow parsing for Khasi has been carried out in the lines of Molina and Pla (2002) where they approached shallow parsing as a tagging problem utilizing the standard HMM approach for parsing without changing the training and tagging process. They have put forward a specialized HMM where adjustments have been made in the training corpus while the training and tagging procedure remains intact. This is carried out by embedding only relevant input information into the model and expanding the chunk tags with supplementary details without affecting the learning and tagging process. The parser is the first of its kind for the language, where the training corpus comprises of 24,194 chunks of noun and verb chunks out of a total of 3,983 sentences and 86,087 tokens. The full details of this process is given in (Tham 2018a). This specialized HMM for Khasi gave an F1 measure of 95.51. This rather optimistic performance is due to the fact that it was tested on gold data that was correctly tagged with POS tags. It was not tested on the output of the POS taggers mentioned earlier. Hence, a deep learning approach using bidirectional gated recurrent (BiGRU) unit was likewise developed for shallow parsing Khasi. The performance of this shallow parser on the same gold data gives an F1 measure of 98.91 (Tham 2020a). However, to get a picture on how its performance will be affected by the performance of the Khasi POS tagger, it was tested on the data tagged by the HMM POS tagger for Khasi and gave an F1 measure of 89.91. This result clearly indicates that improving tagging performance will also improve parsing performance because of the dependency of the parser on POS tagged data. Therefore, the development of the Hybrid POS tagger for Khasi is a much needed improvement in this direction.

5 Conclusion

In this demonstration paper the NLP tools for Khasi and their performances have been presented. Hopefully, with these tools in place it will be rewarding to see their applications in various NLP applications for Khasi.

References

Antonio Molina and Ferran Pla. 2002. Shallow Parsing using Specialized HMMs. *Journal of Machine Learning Research. Introduction to Special Issue on Machine Learning Approaches to Shallow Parsing*, 2:595-613.

Lance Ramshaw and Mitch Marcus. 1995. Text Chunking Using Transformation-Based Learning, in *Proceedings of third Workshop on Very Large Corpora*, pages 82–94.

Medari J. Tham. 2020a. Bidirectional Gated Recurrent Unit For Shallow Parsing. *Indian Journal of Computer Science and Engineering (IJCSE)*, 11(5): 517-521, DOI: 10.21817/indjcse/2020/v11i5/201105167

Medari J. Tham. 2020b. A Hybrid POS Tagger for Khasi, an Under Resourced Language**.** *International Journal of Advanced Computer Science and Applications (IJACSA)*, 11(10):333-342, https://dx.doi.org/10.14569/IJACSA.2020.0111042

Medari J. Tham. 2018a. Khasi Shallow Parser. In *Proceedings of 15th International Conference on Natural Language Processing*. ICON2018, pages 43-49.

Medari J. Tham. 2018b. Challenges and Issues in Developing an Annotated Corpus and HMM POS Tagger for Khasi. In *Proceedings of 15th International Conference on Natural Language Processing*. ICON2018, pages 10-19.