Replication Study of MorphAGram (Based on Tekó Data Set)

Uliana Vedenina
uliana.vedenina@student.uni-tuebingen.de

Abstract

Due to the wide use of morpheme segmentation algorithms in various natural language processing applications, this topic has drawn increased focus in the recent years. One of the publicly available frameworks that addresses such tasks is MorphAGram. Similar to the original study, the current paper aims to evaluate the performance of the model on a small data set for Tekó, a low-resource polysynthetic language, and subsequently compare the obtained results with the described outcomes. The findings signify that on the Tekó dataset the model underperforms using a standard setup.

1 Introduction

Nowadays, morpheme segmentation is a crucial component of many NLP applications. Due to the large amounts of language data and the high cost of the manual annotation for under-resourced languages, the capability to automatically identify distinct morphemes of a given word in both high- and low-resource languages has gained an increased attention in the recent years. One of the frameworks designed to solve morpheme segmentation tasks in an unsupervised and semi-supervised manner is called MorphAGram (Eskander et al., 2020). The original study showed that in both cases the model allows to reach state-of-the-art results on various amounts of data and for different language typologies. The goal for the current project was to replicate the algorithm on the data set of Tekó, the polysynthetic low-resource language, and compare the outcomes with the original results to evaluate the universality of the model across languages, as stated in the original paper.

2 Methodology

The approach proposed by Eskander et al. (2020) for unsupervised and semi-supervised morphological segmentation is built on the Pitman-Yor Adaptor-Grammar Sampler (PYAGS). PYAGS is the variation of probabilistic context-free grammars (PCFGs) combined with an adaptor which integrates the Pitman-Yor process. In turn, this allows to implement existing models for word segmentation and morphological analysis as simple grammars (Johnson et al., 2006). The sampler requires an adapted Context-Free Grammar (CFG) and the encoded data set as inputs for training. In the study, the authors incorporated nine different types of CFG, varying in terms of word modeling, level of abstraction, and boundaries segmentation. To test the Adaptor Grammar in different scenarios for the current task, the algorithm was trained using the Standard, Scholar-seeded, and Cascaded settings. Both Standard and Cascaded setups are suitable for an unsupervised training, although the first one is a fully language-independent configuration, while the Cascaded setting requires a generated set of affixes obtained in the prior learning phase as an input. The Scholar-seeded setup represents a semi-supervised approach which relies on the manually provided set of prefixes and suffixes.

The current study focuses solely on reproducing the algorithm using the Standard setup.

2.1 Data set

The primary idea of the replication attempt considered the use of the original data sets although, due to the number of limitations described in the section 4, it was not achieved. Subsequently, only the Tekó data set was implemented for the analysis. It is subject to the currently unpublished UD_Teko-TuDeT treebank (Vedenina and Gerardi, 2022) version, which includes a set of annotated grammar examples. It is worth mentioning that Tekó is an Amazonian low-resource polysynthetic language which corresponds to the languages type studied in Eskander et al. (2019).

3 Replication Study

The present replication attempt implements the algorithm presented on the MorphAGram GitHub page [1]. The authors provide a description of the necessary steps to preprocess inputs, train the PYAGS, construct a segmentation model, and analyze both the data sets and the model outputs. However, it should be noted that the original code does not include the evaluation algorithm.

As it was already mentioned, the replication of the morpheme segmentation algorithm was not carried out on the original data sets (out of 12, only three data sets for Georgian, Japanese, and Arabic were accessible on the MorphAGram GitHub page). It happened due to several reasons, including the unavailability of test sets for all languages, the absence of gold segmentation for the validation set in Arabic, the mismatch between the file containing the gold segmentation of the validation set and the file containing the unsegmented validation set for Georgian (which resulted in a raised `Value error` during the final evaluation step), and time constraints that prevented the training of the PYAGS for Japanese (having 50,000 data points in the training set).

3.1 Pre-processing

Before the pre-processing of the data sets for the PYAGS, the Tekó corpus was transformed into the *.csv* format by applying the *CoNLL-U_Parser* (Yaseen Khan, 2019), and all the distinct word forms were extracted. Subsequently, the data sets were divided in the 60/20/20 ratio (779, 248, and 251 words respectively).

Similar to the original study, language-specific characters were appended to nine versions of CFGs as terminals.

3.2 Training of PYAGS

The main challenge encountered during the replication process references to the usage of an outdated version of the PYAGS algorithm written in C++ (Johnson et al., 2006). It was released in 2007 and was minorly updated since then. It was discovered that Adaptor Grammars (AG) can not be compiled on MacOS with Apple M2 chip using the default Apple Clang compiler. Running the AG with the brew versions of *gcc* and *c++* also resulted in errors. Consequently, the training phase was carried out on the Cloud9 IDE of the AWS server.

As well as in the original study, the training was conducted on the unsegmented `train+dev` data sets.

4 Results

4.1 Evaluation

Following the training phase, the validation and test sets for Tekó were segmented and subsequently evaluated with **BPR** and **EMMA-2 F1-score** metrics using *morpheval* library [2]. While BPR (Boundary Precision-Recall) calculates the correlation between correctly assigned, incorrectly assigned, and missed boundaries, the EMMA-2 metric is designed for unsupervised morphological tasks and is based on creating morpheme-word graphs for the predicted output and their comparison to the target values (Virpioja et al., 2011).

The most effective models had PrStSu+SM and PrStSu2a+SM grammars as configuration settings. In the first instance, the word is represented as a sequence of prefixes, a stem, and a sequence of suffixes, which are further divided into sub-morphs. In the second instance, the stem and suffixes are represented as a StemSuffixes class.

The outcomes of the performance of the nine models on the validation (transductive approach, wherein the unsegmented word was provided in the training step) and the test sets (inductive approach, wherein the word was not seen by the model previously) are descripted in Tables 1-2.

[1] https://github.com/rnd2110/MorphAGram

[2] https://github.com/svirpioj/morphoeval

Grammar type	BPR F1	Emma-2 F1
PrStSu	0.7326	0.9175
PrStSu+SM	**0.7873**	**0.9382**
PrStSu+Co+SM	0.4588	0.9919
Simple	0.6981	0.9079
Simple+SM	0.7505	0.9007
Morph+SM	0.2473	0.6188
PrStSu2a+SM	**0.7956**	**0.9274**
PrStSu2b+SM	0.7568	0.9246
PrStSu2b+Co+SM	0.7236	0.9114

Table 1: Results on the Tekó validation set (standard setting)

Grammar type	BPR F1	Emma-2 F1
PrStSu	0.5469	0.7259
PrStSu+SM	**0.6267**	**0.766**
PrStSu+Co+SM	0.3375	0.6254
Simple	0.5455	0.727
Simple+SM	0.6073	0.7515
Morph+SM	0.3322	0.6258
PrStSu2a+SM	**0.6267**	**0.7659**
PrStSu2b+SM	0.583	0.7436
PrStSu2b+Co+SM	0.6222	0.7549

Table 2: Results on the Tekó test set (standard setting)

Although the authors claim that the proposed MorphAGram algorithm allows to achieve state-of-the-art outcomes in morpheme segmentation tasks for low-resource polysynthetic languages (Eskander et al., 2019), the training of the PYAGS model using the original parameters, as well as attempts to fine-tune it, resulted in a comparatively subpar performance. This could be due to the fact that the best-performing models presented by the authors have cascaded and scholar-seeded setups, although the model also performs well in the standard setting for these languages. The comparison of the original best-performing models for four low-resource languages based on the PrStSu+SM grammar type (standard setup), together with the results on the Tekó data, is presented in Table 3.

Language	BPR F1	Emma-2 F1
Mexicanero	0.777	0.846
Nahuatl	0.721	0.806
Wixarika	0.768	0.775
Yorem Nokki	0.810	0.863
Tekó	0.627	0.766

Table 3: Model performance on the test set (based on PrStSu+SM grammar)

4.2 Discussion

It should be highlighted that in the original paper the results of the evaluation on all four test sets point to the higher score than on the validation sets, while the unsegmented validation set was seen by the model during the training. It is suggested although not asserted that the high performance outcomes for the test set are explained by the relative morphological simplicity of the included words or the slight variability of morphemes. In contrast, results for Tekó indicate the opposite and, due to the rich morphology of the language, the test set, in particular, may be too complicated for the model to segment. Thus, in several instances, certain n-grams constitute a single morpheme while in other cases, they are separated into multiple morphemes, see patterns in (1)-(2):

(1) dabaʔepatari -> **d-a**-baʔe-pa-tar-i
daepɨdʒiaʔu -> **da**-epɨ-dʒi-aʔu
datɨpʒi -> **da**-tɨpɨ-dʒi
dawarimãʔẽ -> **da**wari-mãʔẽ

(2) ereɲ-> **ere**ɲ
ererek^{w}ar -> **e-re**-rek^{w}ar
erezor -> **ere**-zor

Furthermore, multiple morphemes are present in the test set only once what complicates the segmentation process for the model.

To possibly improve the outcomes' quality, the further study could be based on the implementation of cascaded and scholar-seeded setups and/or the augmentation of the Tekó data set.

5 Conclusion

First of all, the conducted study revealed that for the accurate replication of the MorphAGram algorithm, the materials provided by the authors lack in data. Moreover, the PYAGS model may require modernisation as its last minor update was released in 2013.

The results of the model performance on the Tekó data indicate that the Standard setup is not sufficient to process the language's complex morphology. In the future research, it is aimed to implement two other learning settings to reach enhanced results.

References

Ramy Eskander, Francesca Callejas, Elizabeth Nichols, Judith L Klavans, and Smaranda Muresan. 2020. Morphagram, evaluation and framework for unsupervised morphological segmentation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 7112–7122.

Ramy Eskander, Judith L Klavans, and Smaranda Muresan. 2019. Unsupervised morphological segmentation for low-resource polysynthetic languages. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 189–195.

Mark Johnson, Thomas Griffiths, and Sharon Goldwater. 2006. Adaptor grammars: A framework for specifying compositional nonparametric bayesian models. *Advances in neural information processing systems*, 19.

Uliana Vedenina and Fabrício Gerardi. 2022. UD Teko-Tudet. `https://github.com/UniversalDependencies/UD_Teko-TuDeT`.

Sami Virpioja, Ville T. Turunen, Sebastian Spiegler, Oskar Kohonen, and Mikko Kurimo. 2011. Empirical comparison of evaluation methods for unsupervised learning of morphology. *Traitement Automatique des Langues*, 52(2):45–90.

Muhammad Yaseen Khan. 2019. CoNLL-U_Parser. `https://github.com/MuhammadYaseenKhan/CoNLL-U_Parser`.