International Journal of Research in English
www.englishjournal.net
Online ISSN: 2664-8725; Print ISSN: 2664-8717
Received: 09-01-2019; Accepted: 15-02-2019; Published: 27-02-2019
Volume 1; Issue 1; 2019; Page No. 37-43

Unsupervised morphological segmentation of holophrastic languages

Joginder Singh Ahlawat
Department of English, Hindu College, Sonipat, Haryana, India

DOI: https://doi.org/10.33545/26648717.2019.v1.i1a.312

Abstract
Polysynthetic languages pose a unique challenge for morphological analysis due to their very high morpheme-to-word ratio and very high inflection in word formation. These languages are morphologically rich and highly synthetic, i.e., single words can be composed of many morphemes. In extreme cases, the entire sentence consists of only one single token. Morphological segmentation in such cases becomes a challenging task when we consider the fact that the training data can be extremely scarce for such languages. This paper presents some unsupervised approaches for morphological segmentation of low-resource polysynthetic languages based on Adaptor Grammars (AG) framework.

Keywords: Morphological segmentation, natural language processing, poly- synthetic languages, adaptor grammars

Introduction

In Natural Language Processing (NLP), morphological segmentation is the task of segmenting words into meaning-bearing morphemes. Morphological segmentation is a critical task in NLP. It has many direct or indirect applications in various fields, such as part-of-speech tagging, text classification, machine translation, speech recognition, etc. We present an unsupervised approach for the morphological segmentation of polysynthetic languages based on the Adaptor Grammars framework [1]. The present work showed that the Adaptor Grammars-based approaches outperform other unsupervised methods Morfessor [4] and MorphoChain [5] are considered as state-of-the-artfor this task. Computational morphological segmentation is an active research topic as it forms an essential basis for many natural language processing tasks. Due to the high cost associated with manually labeling data for morphology and the increasing interest in low-resource and endangered languages, where they lack adequate morphologically annotated resources, the need arises to find an unsupervised approach for the task of morphological segmentation, especially for low-resource languages. The main goal is to examine the success of 'Adaptor Grammars'. In the present paper, an approach of unsupervised morphological segmentation has been applied to polysynthetic languages. The proposed approach also deals with synthetically complex (Not simply agglutinative) morphology processes and limited resources. At first, we describe the theoretical aspects of the segmentation, including the results. We begin by defining the problem and highlighting the challenges associated with it. We then describe the proposed solution setup in detail. The last section of this paper deals with the evaluation and results of our model versus the other state-of-the-art baselines. Finally, we deal with the practical aspects of the paper and outline the processes going on behind the scene. We wrap up by discussing and concluding our results in the last section. We also propose the possibility of an extension of this model for Indian languages in the future.

Related Work and Background Knowledge

Adaptor Grammars: Adaptor Grammars are a framework for specifying nonparametric Bayesian models that helps in learning latent tree structures from a corpus of strings. There are two components to an AG model: the base distribution, which is just a PCFG, and the adaptor, which "adapts" the probabilities assigned to individual subtrees under the PCFG model, such that the probability of a subtree under the complete model may be considerably higher than the product of the probabilities of the PCFG rules required to construct it. Although in principle, the adaptor can be any function that maps one distribution onto another. Johnson *et al*. (2007) [1] use a Pitman-Yor Process (PYP) (Pitman and Yor, 1997) as the adaptor because it acts as a caching model. Under a PYP AG model, the posterior probability of a particular subtree will be roughly proportional to the number of times that subtree occurs in the current analysis of the data (with the probability of unseen subtrees being computed under the base PCFG distribution).

Context Free Grammar (CFG)

In formal language theory, a context-free grammar (CFG) is a formal grammar in which every production rule is of the form

$$A \rightarrow \alpha$$

where A is a single nonterminal symbol, and α is a string of terminals and/or nonter- minals (α can be empty). A formal grammar is considered "context free" when its production rules can be applied regardless of the context of a nonterminal.

Probabilistic Context Free Grammar. A Probabilistic Context-Free Grammar (PCFG) is simply a Context-Free Grammar with probabilities assigned to the rules such that the sum of all probabilities for all rules expanding the same non-terminal is equal to one. PCFG parsing takes advantage of probabilities by giving you the most probable parse for a sentence.

Methodology

Objective

Design a statistical machine learning algorithm to split a given word into the surface forms of its smallest meaning-

bearing units, morphemes. The task needs to be per- formed for polysynthetic languages.

For this particular task, we need to segment words from the four Uto-Aztecan languages: Mexicanero (MX), Nahuatl (NH), Wixarika (WX) and Yorem Nokki (YN) [11].

Languages and Datasets

Language Typology and Morphological Analysis: In linguistic typology, the broader gradient of morpheme complexity in a language is isolating/analytic to synthetic to polysynthetic. The more specific gradation is agglutinating to mildly fusional to fusional [2].

Thus, a polysynthetic language can be characterized from agglutinating to fusional.A polysynthetic and agglutinating language has a high number of morphemes per word, with clear boundaries between morphemes. While, a polysynthetic and fusional language also has many morphemes per word, but due to many phonological and other processes, the segmentation boundaries are not clear. Typically, polysynthetic languages demonstrate holophrastic, i.e. the ability of an entire sentence to be expressed as what is considered by native speakers to be just one word [2].

Description of the four Yuto-Aztecan languages (which are considered for analysis):

- "Mexicanero is a Western Peripheral Nahuatl variant, spoken in the Mexican state of Durango by approximately one thousand people [11]."
- "Nahuatl is a large subgroup of the Yuto-Aztecan language family, and, including all of its variants, the most spoken native language in Mexico. The data collected for this work belongs to the Oriental branch spoken by 70 thousand people in Northern Puebla." [11]
- "Wixarika is a language spoken in the states of Jalisco, Nayarit, Durango and Zacatecas in Central West Mexico by approximately fifty thousand people." [11]
- "Yorem Nokki is part of the Taracachita subgroup of the Yuto-Aztecan language family. Its Southern dialect is spoken by close to forty thousand people in the Mexican states of Sinaloa and Sonora, while its Northern dialect has about twenty thousand speakers. In this work, we consider the Southern dialect." [11]

All these four languages are classified as polysynthetic languages [2, 3]. "However, as noted above, there is a gradation of polysynthesis, so the delineation of language types is not clear-cut. For these four languages, the more agglutinative is WX; Leza (2004) has observed 20 morphemes per word for this language." [2]

The morphological analysis of polysynthetic languages thus is challenging due to the root-morpheme complexity. This property of polysynthetic languages makes the task of surface segmentation complex but also relevant for further analysis.

Datasets for Experiment

Morphological data for polysynthetic languages are scarce. Kann *et al*. (2018) proposed a small set of morphologically segmented datasets [11]. These datasets were constructed so that the dataset can contain both segmentable as well as non-segmentable words, which is necessary to ensure that methods can correctly decide against splitting up single morphemes [2].

As we are using the data in an unsupervised way, we have only used the unsegmented data to prepare our training set. There are two different possible approaches for the learning namely the transductive approach and the inductive approach which is explained as follows:

- In transductive mode, any word that needs to be segmented should be in the vocabulary of the training input. The segmentation output is just a lookup to the mapping between training input and the segmentation output.
- In inductive mode, the word need not be present in the training list. In this mode the learner does not process the query words; instead, the segmentation is performed by parsing the words using a PCFG parsing algorithm such as CYK (CYK algorithm is a parsing algorithm for CFGs). In the present work, we have used a transductive mode of learning. Here we include the unsegmented words from the test sets in our training sets. Inductive learning did not improve any performance. We have slightly processed the datasets for clarity purpose. This also helped in satisfying the criteria for the input-output format of our application and for the evaluation of software as mentioned in Kann *et al*. [6].

Using Adaptor Grammar for Polysynthetic Languages-

An Adaptor Grammar is typically composed of a PCFG and an adaptor that adapts the probabilities of individual subtrees. For morphological segmentation, a PCFG is a morphological grammar that specifies word structure, where AGs learn latent tree structures given a list of words [2].

The AGs take as input the vocabulary of the language we want to learn and the grammar rules the sampler has to follow.

- The vocabulary input is a list of unsegmented words.
- The grammar input consists of rules along with the adaptation information

The experiment setup outline is mentioned below

1. Grammars- Nine grammars from Eskander *et al*. (2016) [9] has been used in the present work.
2. Learning Settings- The present work used three learning settings namely Standard, Scholarx-seeded Knowledge and Cascaded as mentioned in Eskander *et al*.

Defining Grammar

The first step of learning morphological segmentation consists of defining the grammar using Adaptor Grammars. As mentioned earlier the present work used nine grammar described by Eskander *et al*. (2016) [2].

Each production rule in our grammar has to be associated with three parameters; θ, a, and b, where θ is the probability of the rule in the generator, while a and b are the parameters of the Pitman-Yor process [10]. If the parameters are not specified, they are sampled by the trainer, or we can also set them to default values before running the learner. Setting a to one means the underlying non-terminal is not adapted and is sampled by the general Pitman-Yor process, while setting a to zero means the adaptor of the non-terminal is a Dirichlet process [10] with the concentration parameter b." [3].

When a non-terminal is adapted, each sub-tree that can be generated using the initial rule of that non-terminal is

considered as a potential rule in the grammar. Otherwise, the non-terminal expands as in a regular PCFG.

Model Training

Inputs

The main inputs to the learner are the grammar and the vocabulary of the language we want to learn the segmentation for. It is to be noted that the standard grammar is a comprehensive set of rules and is entirely independent of the language. Also, note that as we are using the transductive learning method, the test set unsegmented words should be included in the train set.

```
1 1 Word --> Prefix Stem Suffix

Prefix --> ^^^
Prefix --> ^^^ PrefixMorphs

1 1 PrefixMorphs --> PrefixMorph PrefixMorphs
1 1 PrefixMorphs --> PrefixMorph
PrefixMorph --> SubMorphs

Stem --> SubMorphs

Suffix --> $$$
Suffix --> SuffixMorphs $$$

1 1 SuffixMorphs --> SuffixMorph SuffixMorphs
1 1 SuffixMorphs --> SuffixMorph
SuffixMorph --> SubMorphs

1 1 SubMorphs --> SubMorph SubMorphs
1 1 SubMorphs --> SubMorph
SubMorph --> Chars

1 1 Chars --> Char
1 1 Chars --> Char Chars
```

Fig 1: The standard PrStSu+SM grammar (note the simple recursive rules of the grammar)

We use A A A and $$$ to indicate the beginning and end of words respectively. The input grammar to the AGs should follow the format similar to one shown in Fig. 1.

Learning Settings

Escander *et al*., (2016) [9] define three learning settings: Standard, Scholar-seeded and Cascaded. The detail description is mentioned below..

- **Standard:** The standard setting is language independent. The grammar does not have any language-specific rules, and the learning is fully unsupervised. Fig. 1 shows the input of PrStSu+SM grammar in the standard mode.
- **Scholar-seeded:** When some linguistic knowledge is available (such as a list of morphemes) this knowledge can be seeded into the grammar trees as additional production rules, allowing for a semi-supervised learning setup [3].
- **Cascaded:** The cascaded setting approximates the effect of a scholar-seeded setting in a language-independent setup. This is done by first obtaining a list of morphemes from a segmentation model that is trained on a high-precision grammar, and then seeding those morphemes into another grammar. In this setup, both the standard grammar and the segmentation output of another grammar are provided. The system then extracts the top morphemes, typically affixes, from the segmentation output and seeds them into the standard grammar prior to running the learner [3]."

Also, note that the grammars are not ready to be used as input for the Adaptor Grammars framework. The Char (in grammar rules) needs to be assigned to every letter terminal in the alphabet of the language for which we want to learn the segmentation. However, this process can be automated by picking up the characters from the training set and defining the grammar then.

Evaluation and Results

Here, we evaluate our morphological segmentation setup, qualitatively and analytical- ly. We first describe the evaluation setup, then we describe the evaluation metrics and the baselines. We conclude by comparing our results to the state-of-the-art baselines.

Evaluation Setup: LIMS (Language-Independent Morphological Segmented) is a system that provides the best on average results for morphological segmentation using Adaptor Grammars. It uses the cascade, starting with PrStSu2b+Co+SM grammar, and inserting the obtained affixes into PrStSu+SM and running this modified PrStSu+SM [9].

- For our model, we conduct the evaluation using the LIMS setup, as proposed by Eskander *et al*., (2016) [9]. We found that the LIMS setup gave the best performance on average for all these languages. LIMS is a cascaded setup with a high-precision grammar (such as PrStSu2b+Co+SM) as the input grammar for the first run and a PrStSu+SM grammar as the second grammar input to the AG. We only report the results for this setup, as this provides the best results for our model.
- We conduct the evaluation in a transductive learning scenario, where the unsegmented test words are included in our training set.
- No annealing is used as it does not improve the results. All the parameters to the Adaptor Grammars are automatically inferred (We do not specify any default value for any of the hyperparameters).
- We compute the results as the average of five runs.

Evaluation Metrics: The performance of the present work has been analyzed using two metrics: Boundary Precision and Recall (BPR) and Evaluation Metric for Morphological Analysis - 2 (EMMA-2) [12].

BPR is the classical evaluation method for morphological segmentation, where the boundaries in the proposed segmentation are compared to the boundaries in the reference. In contrast, EMMA-2 is based on matching the morphemes, and is a variation of EMMA (Spiegler and Monson, 2010) [12]. In EMMA, each proposed morpheme is matched to each morpheme in the gold segmentation through one-to-one mappings. However, EMMA-2 allows for shorter computation times as it replaces the one-to-one assignment problem in EMMA by too many-to-one assignment problems, where two or more proposed morphemes can be mapped to one reference morpheme.

EMMA-2 also results in higher precision and recall as it tolerates failing to join two allomorphs or distinguishing between identical syncretic morphemes [3].

Baselines: We evaluate our system versus two state-of-the-art baselines: Morfessor [4] and MorphoChain [5].

- Morfessor is a commonly used framework for unsupervised and semi-supervised morphological segmentation and is publicly available free.
- MorphoChain is another publicly available system for unsupervised morphological segmentation.

System Performance

Table 1 reports the performance of our system and Table 2 reports our performance compared to Morfessor and MorphoChain based on the re- spective F1 - scores for the four polysynthetic languages when tested on Test sets. The results by our system are under the AG-LIMS column.

Table 1: Result on Test sets using the AG based LIMS model (AG-LIMS)

Language		BP R			EMMA-2	
	EMMA Precision	Recall	F1 Score	Precision	Rec all	F1 Score
Mexicanero	67.8	77.3	71.8	85.6	81.8	82.4
Nahuatl	65.1	72.3	67.4	81.1	78.0	78.6
Wixarika	84.3	66.8	73.3	87.2	65.5	72.1
Yorem Nokki	84.0	75.9	78.5	93.1	80.8	84.8

The high precision and low recall for WX suggests that our model tends to under- segment the words of this language. This can be attributed to the high morphemeto- word ratio for WX and existence of many single letter morphemes in the language. Our model fails to segment these single letter morphemes.

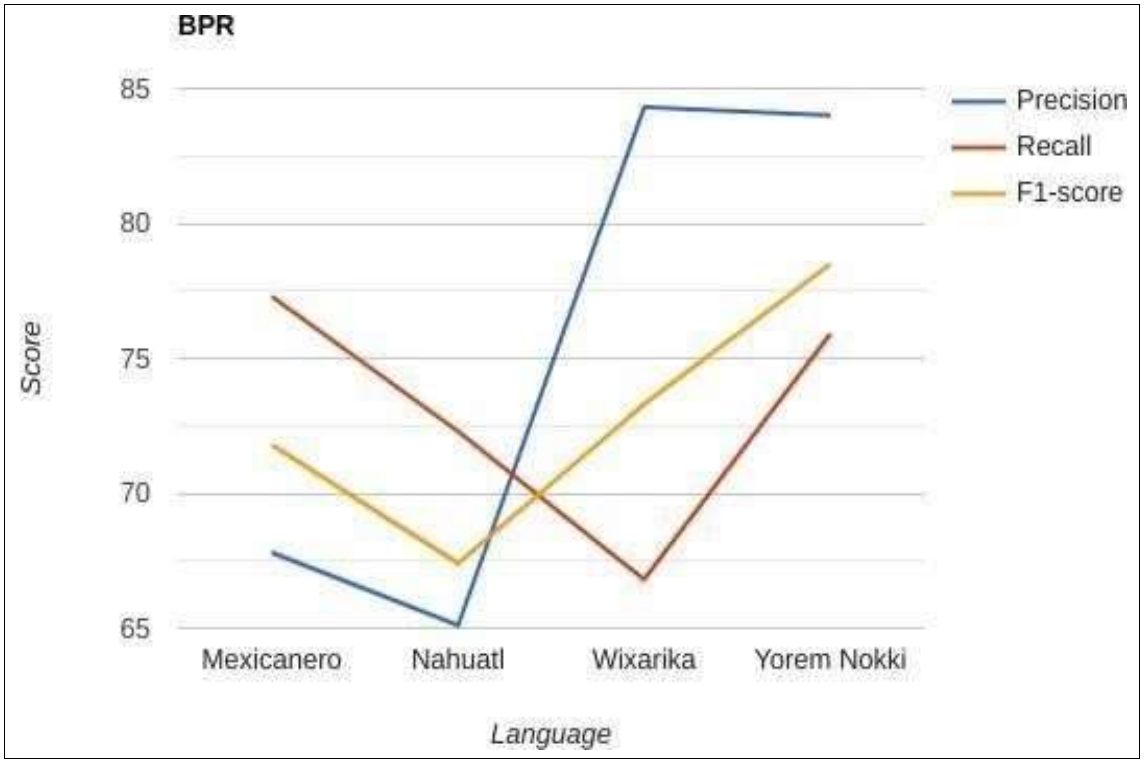

Fig 2: BPR scores of each language (from Table 1)

As shown in Table 1, we obtain better scores on both metrics, compared to our baselines, on all four languages. "It is worth noting that models that tend to under segment achieve significantly better EMMA-2 scores as opposed to the BPR ones, which is due to the one-to-many mappings in EMMA-2. The system ranking may differ depending on the evaluation metric.

Table 2: Result on test sets using the AG based LIMS model (AG-LIMS) for each language compared to two baselines: Morfessor and MorphoChain. The results are reported on both the BPR and EMMA-2 F1-scores.

Language		BP R			EMMA-2	
	Morfessor	Morpho Chain	AG LI MS	Morfessor	Morpho Chain	AG LI MS
Mexicaner o	70.5	64.3	71.8	79.6	77.4	82.4
Nahuatl	61.2	55.9	67.4	73.4	74.8	78.6
Wixarika	72.9	50.3	73.3	71.7	62.3	72.1
Yorem Nokki	71.7	58.0	78.5	79.0	77.6	84.4
Average	69.1	57.1	72.8	75.9	73.0	79.5

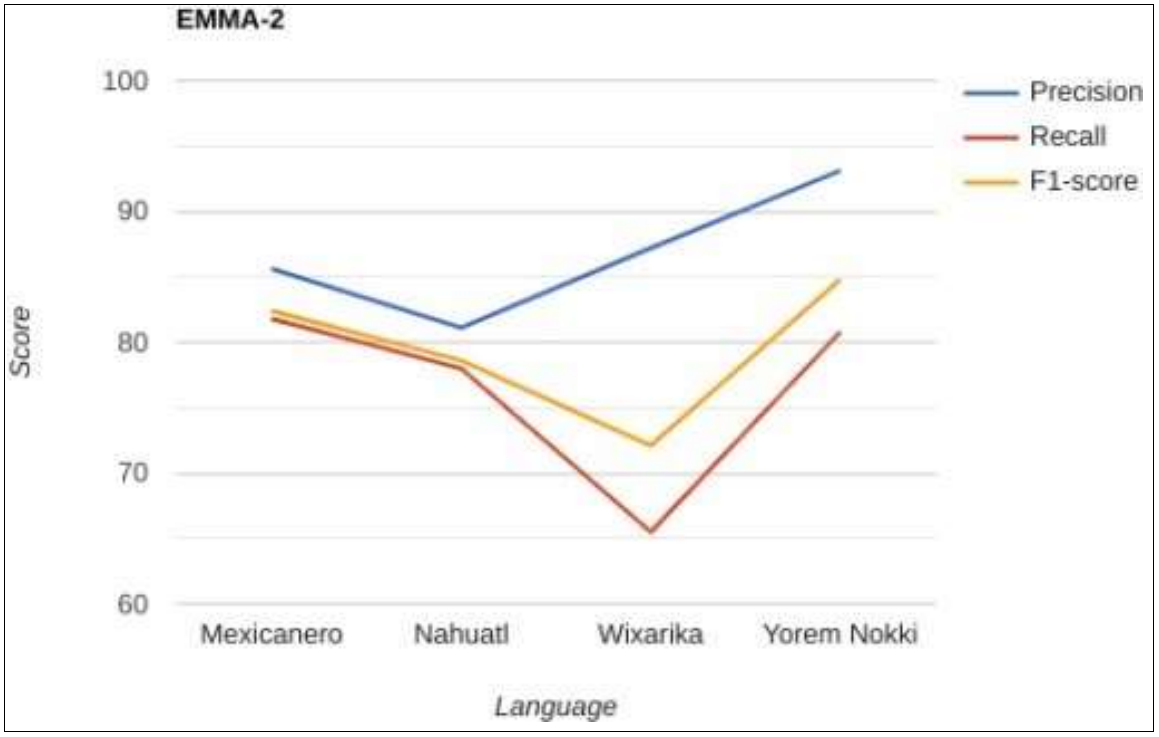

Fig 3: EMMA-2 scores of each language (from Table 1)

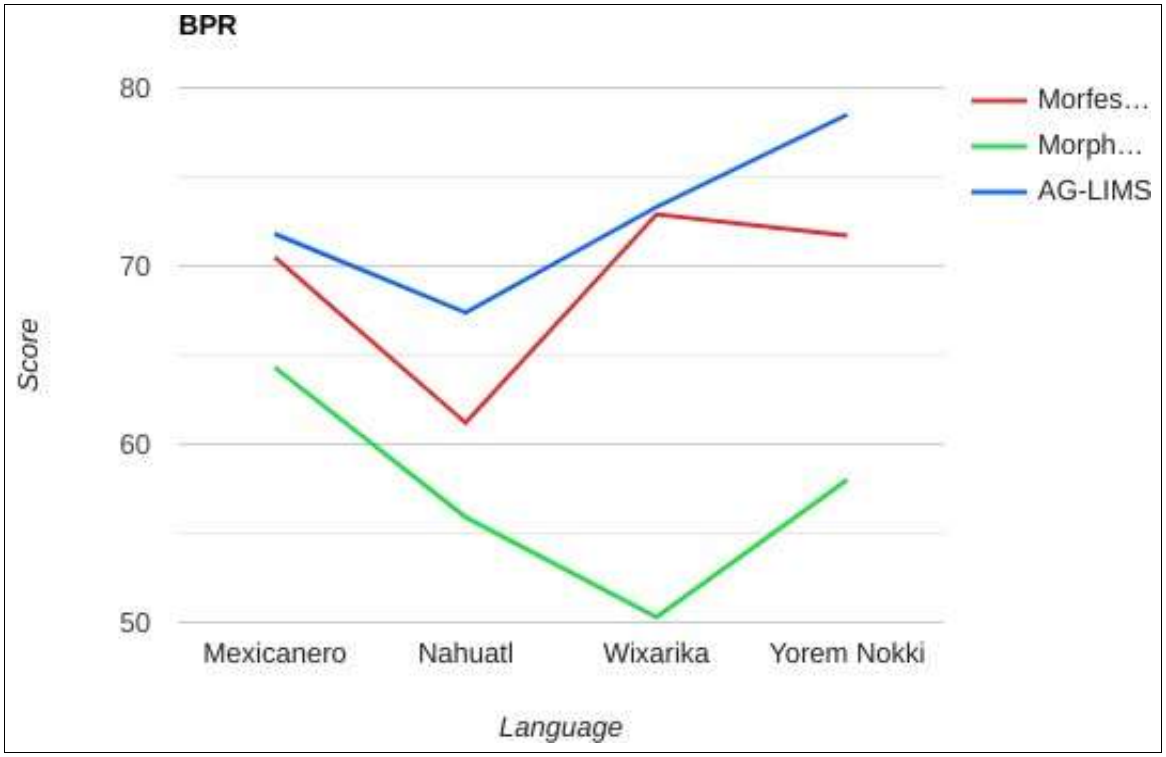

Fig 4: Comparison of BPR F1-scores of the setups, across all languages (from Table 2)

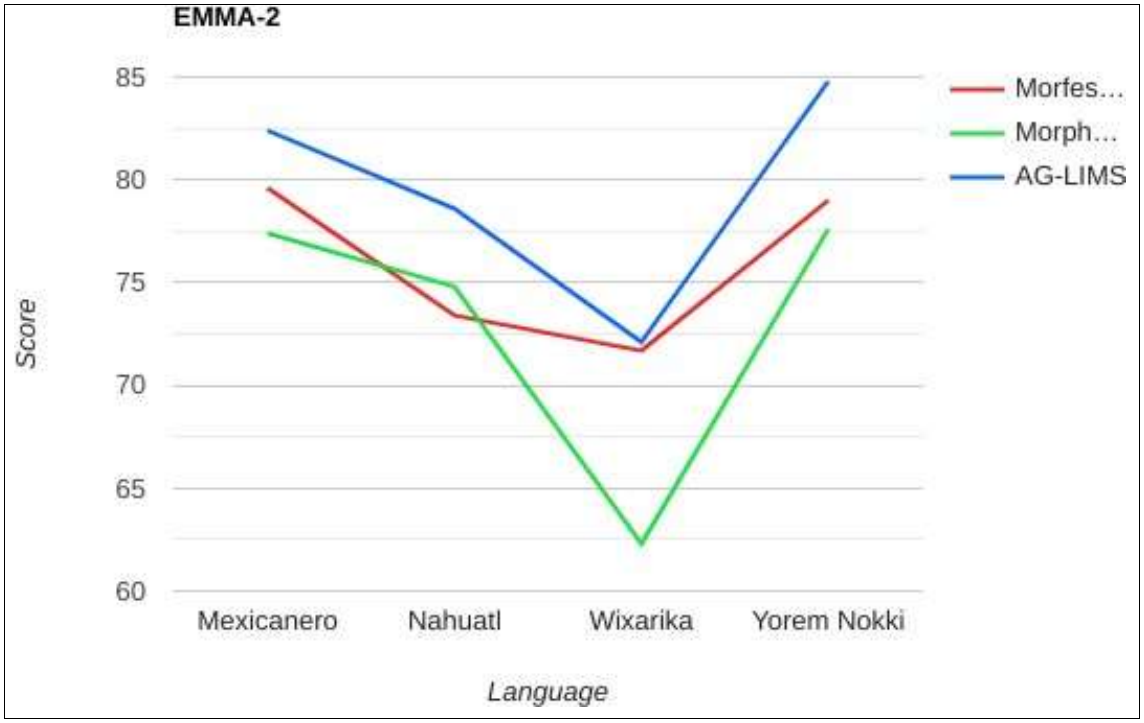

Fig 5: Comparison of EMMA-2 F1-scores of the setups, across all languages (from Table 2)

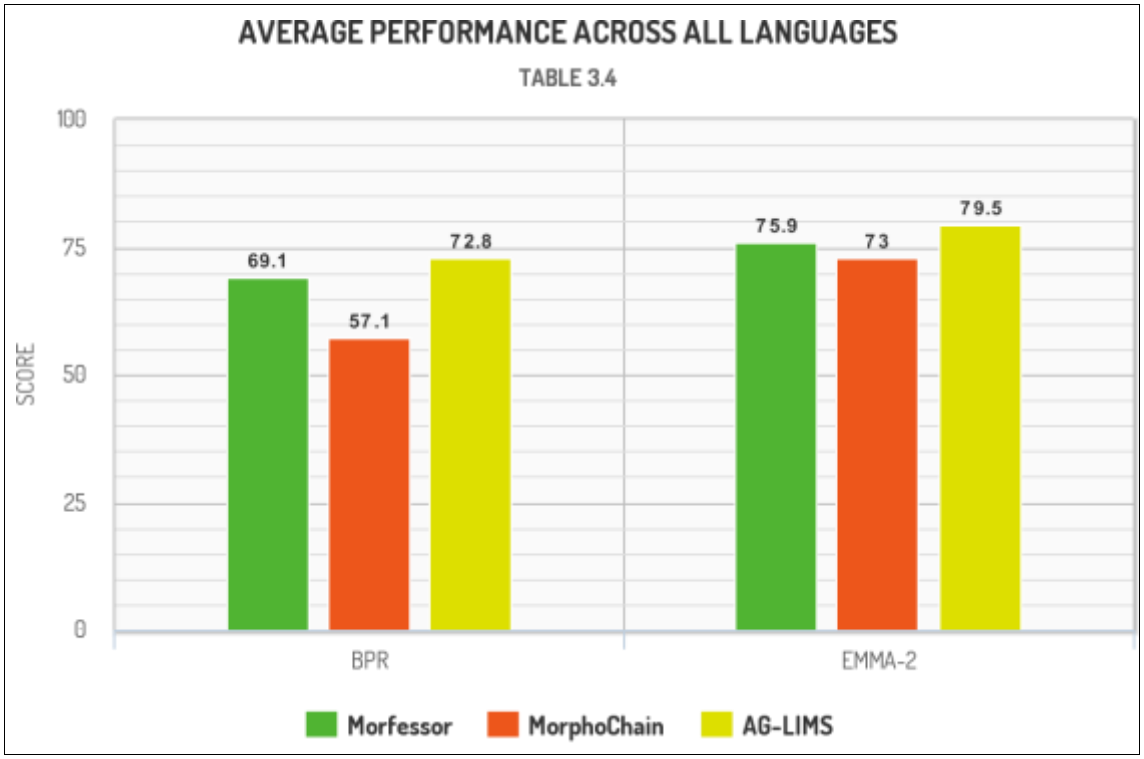

Fig 6: Comparison of the average F1 score of the setups (BPR scores on the left side and EMMA-2 scores on the right side of the graph)

Conclusion

As can be seen from Table 2, Fig. 4 and Fig. 5, we constantly obtain better results versus Morfessor and MorphoChain, for each language when tested using BPR and EMMA-2 metrics. The AG-LIMS setup achieves the best result on all languages in an unsupervised manner. The AG-LIMS setup achieves absolute average F1-score increases of 3.7% and 15.7% over Morfessor and MorphoChain respectively using the BPR metric. The performance analysis shows that Adaptor Grammars-based system is able to generalize and learn well from a small amount of data.

Result Analysis and Scope of Future Work

The unsupervised approaches using Adaptor Grammar showed promising results in the area of morphological segmentation of low-resource polysynthetic languages.

- The proposed setup showed state-of-the-art results in all languages on both metrics.

- The proposed approach showed that even with a very small amount of unsegmented data (training data) the setup is able to produce promising results.

In the future, this setup can be used for morphological segmentation of some Indian agglutinative languages (e.g., Dravidian languages). As a large morphological dataset is not available for most Indian languages, hence the described setup is promising for this task.

References

1. Johnson M, Griffiths TL, Goldwater S. Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models. In: Scholkopf B, Platt JC, Hoffman T, editors. Advances in Neural Information Processing Systems 19 (NIPS 2006). Vol. 19. MIT Press; c2007. p. 641-648.
2. Eskander R, Klavans J, Muresan S. Unsupervised morphological segmentation for low-resource polysynthetic languages. In: Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology. Florence, Italy: Association for Computational Linguistics; c2019 Aug. p. 189-195. [Online]. Available: https://www.aclweb.org/anthology/W19-4222
3. Eskander R, Callejas F, Nichols E, Klavans J, Muresan S. MorphAGram, evaluation and framework for unsupervised morphological segmentation. In: Proceedings of the 12th Language Resources and Evaluation Conference. Marseille, France: European Language Resources Association; c2020 May. p. 7112-7122. [Online]. Available: https://www.aclweb.org/anthology/2020.lrec-1.879
4. Creutz M, Lagus K. Unsupervised models for morpheme segmentation and morphology learning. ACM Transactions on Speech and Language Processing. 2007;4(1):1-34.
5. Narasimhan K, Barzilay R, Jaakkola TS. An unsupervised method for uncovering morphological chains. Transactions of the Association for Computational Linguistics. 2015;3:157-167. [Online]. Available: http://arxiv.org/abs/1503.02335
6. Kann K, Mager Hois JM, Meza-Ruiz IV, Schütze H. Fortification of neural morphological segmentation models for polysynthetic minimal-resource languages. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). New Orleans, Louisiana: Association for Computational Linguistics; c2018 Jun. p. 47-57. [Online]. Available: https://www.aclweb.org/anthology/N18-1005
7. Sirts K, Goldwater S. Minimally-supervised morphological segmentation using adaptor grammars. Transactions of the Association for Computational Linguistics. 2013;1:255-266. [Online]. Available: https://www.aclweb.org/anthology/Q13-1021
8. Mager M, Gutierrez-Vazquez X, Sierra G, Meza-Ruiz I. Challenges of language technologies for the indigenous languages of the Americas. In: Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe, New Mexico, USA: Association for Computational Linguistics; c2018 Aug. p. 55-69. [Online]. Available: https://www.aclweb.org/anthology/C18-1006
9. Eskander R, Rambow O, Yang T. Extending the use of adaptor grammars for unsupervised morphological segmentation of unseen languages. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. Osaka, Japan: The COLING 2016 Organizing Committee; c2016 Dec. p. 900-910. [Online]. Available: https://www.aclweb.org/anthology/C16-1086
10. Pitman J, Yor M. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. Annals of Probability. 1997;25(2):855-900. [Online]. Available: https://doi.org/10.1214/aop/1024404422
11. Ishwaran H, James LF. Generalized weighted Chinese restaurant processes for species sampling mixture models. Statistica Sinica. 2003;13:1211-1235.
12. Virpioja S, Turunen VT, Spiegler S, Kohonen O, Kurimo M. Empirical comparison of evaluation methods for unsupervised learning of morphology. Traitement Automatique des Langues. 2011;52(2):45-90.