Minimally-Supervised Morphological Segmentation using Adaptor Grammars with Linguistic Priors

Ramy Eskander[1], **Cass Lowry**[2], **Sujay Khandagale**[1], **Francesca Callejas**[1], **Judith Klavans**[3], **Maria Polinsky**[3], **Smaranda Muresan**[1]
[1]Columbia University, {rnd2110,sk4746,ffc2108,smara}@columbia.edu
[2]The Graduate Center, City University of New York, clowry@gradcenter.cuny.edu
[3]University of Maryland, {jklavans,polinsky}@umd.edu

Abstract

With the increasing interest in low-resource languages, unsupervised morphological segmentation has become an active area of research, where approaches based on Adaptor Grammars achieve state-of-the-art results. We demonstrate the power of harnessing linguistic knowledge as priors within Adaptor Grammars in a minimally-supervised learning fashion. We introduce two types of priors: 1) grammar definition, where we design language-specific grammars; and 2) linguist-provided affixes, collected by an expert in the language and seeded into the grammars. We use Japanese and Georgian as respective case studies for the two types of priors and introduce new datasets for these languages, with gold morphological segmentation for evaluation. We show that the use of priors results in error reductions of 8.9 % and 34.2 %, respectively, over the equivalent state-of-the-art unsupervised system.

1. Introduction

Morphological segmentation is an essential subtask in many natural language processing (NLP) applications, especially in the case of morphologically complex languages. With the need to develop NLP tools for low-resource languages, unsupervised morphological segmentation has been receiving increasing interest over the last two decades (Goldsmith, 2001; Creutz and Lagus, 2007a; Poon et al., 2009; Sirts and Goldwater, 2013; Botha and Blunsom, 2013; Narasimhan et al., 2014; Eskander et al., 2016, 2018, 2019).

In this work, we show how linguistic priors effectively boost morphological-segmentation performance in a minimally-supervised manner that does not require segmented words for training. We integrate our priors within Adaptor Grammars (Johnson et al., 2007), a type of nonparametric Bayesian models that generalize Probabilistic Context-Free Grammars (PCFGs). Adaptor Grammars have proved successful for unsupervised morphological segmentation, achieving state-of-the-art results across a variety of typologically diverse languages (Eskander et al., 2020).

We introduce two types of linguistic priors: 1) grammar definition, where we design a language-specific grammar that is tailored for the language of interest by modeling specific morphological phenomena, and 2) linguist-provided affixes, where an expert in the underlying language compiles a list of carefully selected affixes and seeds it into the grammars prior to training the segmentation model. We use Japanese and Georgian as case studies for priors 1 and 2, respectively. As our goal is to develop a robust approach that benefits low-resource and/or endangered languages of high morphological complexity, we use Japanese and Georgian in a low-resource setting where we do not have access to morphologically segmented data for training but have access to linguistic information such as word structure and affixes.

We show that using linguistic priors in a minimally-supervised setting leads to a significant improvement in performance over the equivalent state-of-the-art unsupervised system. We also present two morphologically segmented datasets for Japanese and Georgian that we use as our gold standard and that can be utilized in other morphology tasks.[1]

2. Linguistic Priors

We utilize MorphAGram (Eskander et al., 2020)[2], an open-source morphological-segmentation framework that is based on Adaptor Grammars (AGs) (Johnson et al., 2007). AGs have proved successful for unsupervised and

[1]The training and evaluation datasets, linguistic priors and models for both Japanese and Georgian are available at https://github.com/rnd2110/MorphAGram/data.

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

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 3969–3974

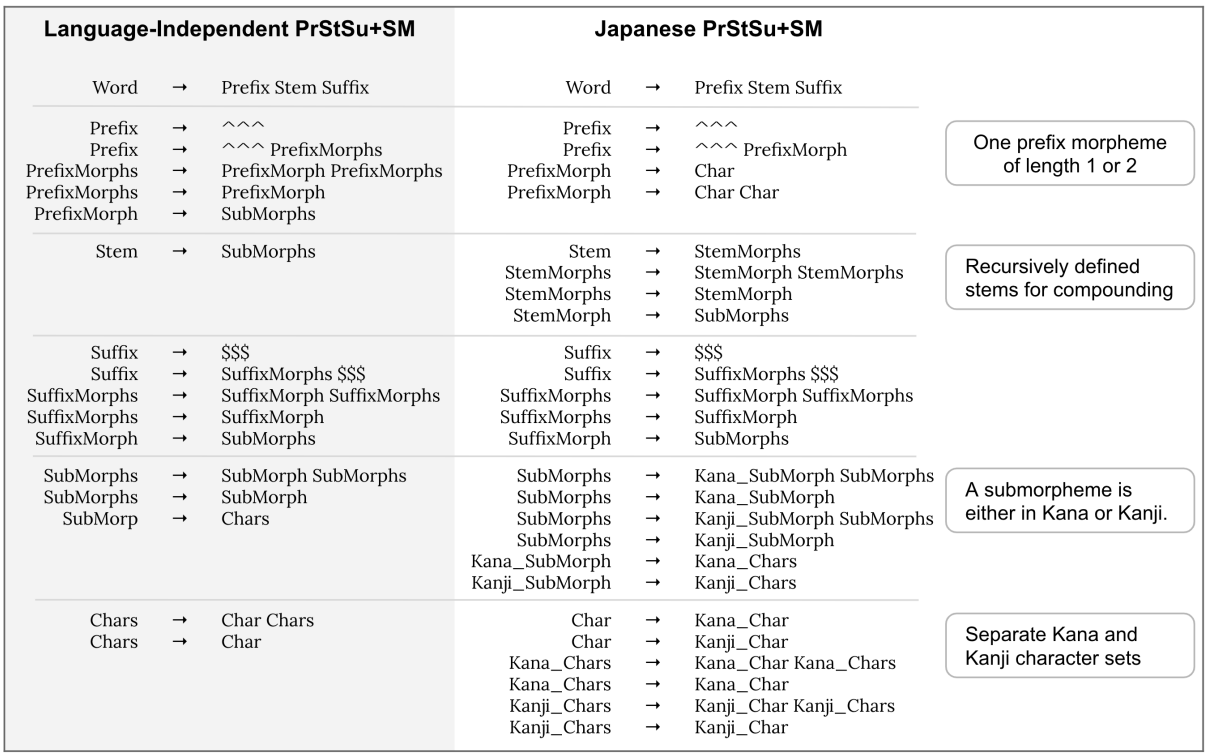

Figure 1: Language-independent PrStSu+SM grammar (left side) vs. its Japanese cognate (right side)

minimally-supervised morphological segmentation, outperforming the competing discriminative models (Sirts and Goldwater, 2013; Eskander et al., 2019, 2020).

Adaptor Grammars are non-parametric Bayesian models that are composed of two main components: 1) a Probabilistic Context-Free Grammar (PCFG) whose definition relies on the underlying task (in the case of morphological segmentation, a PCFG models word structure); and 2) an adaptor that is based on the Pitman-Yor process (Pitman, 1995). The adaptor keeps the posterior probability of a subtree proportional to the number of times that subtree is utilized to parse the input data and manages the caching of the subtrees. The learning process is Markov Chain Monte Carlo sampling (MCMC) (Andrieu et al., 2003) that does the inference of the PCFG probabilities and the hyperparameters of the model.

Eskander et al. (2016) define a set of language-independent grammars and three learning settings for Adaptor Grammars: 1) *Standard*, fully unsupervised; 2) *Scholar-Seeded*, minimally-supervised by manually seeding affixes into the grammar prior to training the segmentation model, and 3) *Cascaded*, fully unsupervised by approximating the *Scholar-Seeded* setting using automatically generated affixes from an initial round of learning. We next present two ways of including linguistic priors in Adaptor Grammars: 1) defining a language-specific grammar; and 2) using linguist-provided affixes in the Scholar-Seeded learning setup.

2.1. Linguistic Priors as Grammar Definition

Eskander et al. (2016) define language-independent grammars that model the word as a sequence of generic morphemes or as a sequence of prefixes, stem and suffixes. We consider their *PrStSu+SM* grammar in the current study as it is the grammar that performed best on average across different languages. This language-independent definition of the grammar is depicted on the left side of Figure 1, where the word is modeled as a prefix *Pr*, a stem *St* and a suffix *Su*, and both the prefix and suffix are recursively defined in order to model compounding in affixes, while a morpheme is composed of smaller units, submorphemes *SM*, representing sequences of characters.

While this grammar is intended to be generic and to describe word structure in any language, we hypothesize that a definition that imposes language-specific constraints would be more efficient. Therefore, we define a grammar for Japanese, where we use characteristics that are specific to Japane-

se word structure as language priors. Our tailored grammar definition for Japanese is shown on the right side of Figure 1, where we impose the following specifications:

- A word has a maximum of one one-character or two-character prefix morphemes.
- A stem is recursively defined as a sequence of morphemes in order to allow for stem compounding.
- Characters are separated into two groups, Kana (Japanese syllabaries) and Kanji (adapted Chinese characters).
- A submorpheme represents a sequence of characters that is either in Kana or Kanji.

2.2. Linguistic Priors as Linguist-Provided Affixes

Similar to the *Scholar-Seeded* setting, we compile a list of affixes and seed it into the grammar trees before learning the segmentation model. However, unlike Eskander et al. (2016), where the affixes are collected from online resources by someone who may have never studied the language of interest, in this study we use affixes that are carefully compiled by an expert linguist who specializes in Georgian, resulting in more accurate linguistic priors. With that goal in mind, a total of 119 affixes are collected from the leading reference grammar book (Aronson, 1990).

3. Evaluation Data

We annotate two datasets with morphological segmentation that we use as the gold standard to evaluate our segmentation models for Japanese and Georgian. Both datasets are composed of 1,000 words that are randomly sampled from the most frequent 50,000 words in Wikipedia and segmented into their basic morphemes[3], similar to the data of the Morpho Challenge shared task [4]. Table 1 lists segmentation examples for both languages.

The Japanese gold segmentation was created by a native-speaker linguist. For Georgian, which has highly complex morphology, we started with the gold-standard dataset of 1000 words introduced by Eskander et al. (2020), which was built by an untrained native speaker and contained only one possible segmentation per word. An expert in Georgian then corrected 193 examples in the data and further annotated 116 words for two possible alternative segmentations. In addition, the expert coded each word based on its syntactic category: verbs (359), nouns (475), numerals (44) and other (122).

Japanese Word	Segmentation
いました	い + ま + した
勉強して	勉強 + し + て
始められません	始め + られ + ま + せん

Georgian Category	Word	Segmentation
Verb	იქნებაო	ი + ქნ + ებ + ა + ო იქნებ + ა + ო
Noun	თვითთვრინავი	თვი + თ + თვრინ + ავ + ი თვითთვრინავ + ი
Numeral	თოთხმეთი	თ + ოთხ + მეთ + ი თოთხმეთ + ი
Other	ვიხი	ვი + ხ + ი ვიხ + ი

Table 1: Japanese and Georgian segmentation examples

4. Evaluation and Results

4.1. Experimental Setup

We evaluate our morphological-segmentation models for Japanese in the *Standard (STD)* and *Cascaded (CAS)*[5] settings, both with generic and language-specific *(LS)* grammar definitions. For Georgian, we evaluate our morphological-segmentation models in the *Standard (STD)*, *Cascaded (CAS)* and *Scholar-Seeded (SS)* settings, in addition to the proposed *Scholar-Seeded* setting with linguist-provided affixes (*SS-Ling*).

We perform the evaluation in a transductive manner, where the unsegmented words in the gold standard are part of the training sets; this is common in evaluating unsupervised and minimally-supervised morphological segmentation (Poon et al., 2009; Sirts and Goldwater, 2013; Narasimhan et al., 2014; Eskander et al., 2016, 2019, 2020). For the metrics, we use Boundary Precision and Recall (BPR) and EMMA-2 (Virpioja et al., 2011). BPR is the classical metric for evaluating morphological segmentation; it compares the boundaries in the proposed segmentation to those in the reference. EMMA-2

[3]The Georgian dataset contains five non-words and three phonetic spellings of English character names.

[4]http://morpho.aalto.fi/events/morphochallenge/

[5]For the Cascaded setup, we use the high-precision grammar *PrStSu2a+SM* defined by Eskander et al. (2016) as the base grammar.

Language	Setting	BPR			EMMA-2		
		Prec.	Recall	F1-Score	Prec.	Recall	F1-Score
Japanese	*Morfessor*	81.4	77.2	79.3	91.5	80.1	85.4
	AG STD	81.7	77.4	79.5	91.0	81.8	86.1
	AG CAS	80.9	78.2	79.5	90.8	82.0	86.2
	AG STD-LS	83.5	79.3	**81.3**	92.0	82.5	**87.0**
	AG CAS-LS	82.8	79.3	81.0	91.1	82.6	86.6
Georgian	*Morfessor*	79.2	54.6	64.6	88.5	56.1	68.7
	AG STD	81.8	69.0	74.9	87.8	65.5	75.0
	AG CAS	83.5	70.4	76.4	88.6	67.2	76.4
	AG SS	84.5	69.1	76.0	89.3	65.2	75.4
	AG SS-Ling	84.6	82.4	**83.5**	87.6	78.2	**82.6**

Table 2: Morphological-segmentation performance for Japanese and Georgian using the BPR and EMMA-2 metrics. The best F1-score per language-metric pair is in **bold**. *AG = Adaptor Grammars*. *STD = Standard*. *CAS = Cascaded*. *STD-LS = Standard* with a language-specific grammar. *CAS-LS = Cascaded* with a language-specific grammar. *SS = Scholar-Seeded*. *SS-Ling = Scholar-Seeded* with linguist-provided affixes

Category	BPR						EMMA-2					
	AG SS			AG SS-Ling			AG SS			AG SS-Ling		
	Prec.	Recall	F1-Score	Prec.	Recall	F1-Score	Prec.	Recall	F1-Score	Prec.	Recall	F1-Score
Noun	74.4	79.6	76.9	74.6	90.4	81.8	87.4	78.7	82.8	84.4	86.8	85.6
Verb	95.8	50.5	66.1	96.6	68.9	80.4	96.4	49.2	65.1	96.1	69.6	80.7
Numeral	93.9	74.1	82.8	87.9	84.8	86.3	87.3	65.5	74.8	81.6	66.0	73.0
Other	87.0	81.6	84.2	86.7	90.3	88.4	92.4	79.0	85.2	92.0	85.8	88.8

Table 3: Category-wise morphological-segmentation performance for Georgian using the BPR and EMMA-2 metrics. *AG = Adaptor Grammars*. *SS = Scholar-Seeded*. *SS-Ling = Scholar-Seeded* with linguist-provided affixes.

is based on matching the morphemes in the proposed segmentation to those in the reference in a many-to-one assignment setup.

We evaluate our system versus two state-of-the-art unsupervised baselines: MorphAGram without the use of linguistic priors and Morfessor (Virpioja et al., 2013) [6]. Morfessor is a commonly-used framework for unsupervised morphological segmentation. It is based on an HMM model that relies on the Minimum Description Length (MDL) concept for deriving the optimal segmentation (Creutz and Lagus, 2007b). Since our approach does not assume access to manually annotated segmentation, it is not directly comparable to semi-supervised approaches that rely on such annotations (Ruokolainen et al., 2014; Kann et al., 2018). Finally, we report all the Adaptor-Grammar results as the average over three runs of different randomization parameters.

4.2. System Performance

Table 2 reports the overall performance of our models for both Japanese and Georgian, while Table 3 shows the results per part-of-speech category for Georgian.

For Japanese, the use of a language-specific grammar definition improves both precision and recall, resulting in BPR F1-score error reductions of 8.9 % and 7.1 % over the generic *Standard* and *Cascaded* settings, respectively, and a BPR F1-score error reduction of 9.8 % over Morfessor.

For Georgian, the use of linguist-provided seeded affixes improves both precision and recall, where the recall significantly increases by absolute 13.3 % over using an affix list of lower quality. In addition, the proposed linguistic priors result in BPR F1-score error reductions of 34.2 %, 30.0 % and 31.1 % over the *Standard*, *Cascaded* and regular *Scholar-Seeded* settings, respectively, and a BPR F1-score error reduction of 53.3 % over Morfessor. Analysing results per category, verbs and nouns receive the biggest F1-score improvements of absolute 14.3 % and 4.9 %, respectively, with the use of linguist-provided affixes.

A similar pattern of results is found with EMMA-2. Finally, all the improvements due to the use of linguistic priors are statistically significant ($P < 0.01$) on both metrics.

4.3. Error Analysis

Table 4 lists some examples of correctly and incorrectly segmented words by our Japanese and

[6] https://morfessor.readthedocs.io/en/latest/

	Word	Gold Segmentation	*STD* Segmentation	*STD-LS* Segmentation
Japanese	お電話	お + 電話	お + 電話	お + 電話
	ご親族	ご + 親 + 族	ご + 親 + 族	ご + 親 + 族
	登られませ	登 + られ + ま + せん	登ら + れ + ません	登 + られ + ません
	比べなかったら	比べ + な + かった + ら	比べ + なかった + ら	比べ + なかった + ら

	Word	Gold Segmentation	*STD* Segmentation	*SS* Segmentation	*SS-Ling* Segmentation
Georgian	ლურჯი	ლურჯ + ი	ლურჯ + ი	ლურჯ + ი	ლურჯ + ი
	რვის	რვ + ის	რ + ვის	რვ + ის	რვ + ის
	გამოვა	გა მო ვ ა	გამო ვ ა	გა მო ვ ა	გამოვ ა
	შაური	შაურ + ი	შა + ური	შა + ურ + ი	შაურ + ი

Table 4: Examples of output segmentations for Japanese and Georgian. *STD* = *Standard*. *STD-LS* = *Standard* with a language-specific grammar. *SS* = *Scholar-Seeded*. *SS-Ling* = *Scholar-Seeded* with linguist-provided affixes. Incorrect morphemes are marked in red.

Georgian segmentation models. We discuss the most prominent observations below.

Japanese: Both the *STD* and *STD-LS* models perform well on prefix segmentation, achieving F1-scores of more than 90 % in the detection of several one-character prefixes, such as お and ご. However, *STD-LS* outperforms its language-independent counterpart in the detection of stems, where compounding is explicitly modeled. For instance, *STD* and *STD-LS* achieve F1-scores of 15.8 % and 98.6 %, respectively, in the detection of the common stem られ (*be*). On the other hand, when either model consistently fails to detect a specific morpheme, the other model fails as well. For example, neither model can detect the morphemes せん and かった.

Georgian: *SS-Ling* outperforms both *STD* and *SS* at discovering the top most frequent one-letter morphemes, such as ი, ა, ს, ე, მ, ო and ვ, achieving an average F1-score of 76.0 %, compared to 57.7 % and 57.3 % by *STD* and *SS*, respectively. In addition, *SS* and *STD* suffer lower precision as they tend to oversegment the morphemes represented by a single letter. Similarly, *SS-Ling* can recognize the most frequent two-letter morphemes, namely ებ and და, with absolute increases in precision of 59.0 % and 62.0 % over *STD* and *SS*, respectively; both morphemes are explicitly seeded into the *SS-Ling* grammar prior to training the model.

5. Conclusion and Future Work

We proposed two types of linguistic priors for minimally-supervised morphological segmentation using Adaptor Grammars. The first prior is in the form of defining a language-specific grammar, while the second relies on compiling a list of linguist-provided affixes and seeding it into the grammars. Our approaches result in error reductions of 8.9 %, for Japanese, and 34.2 %, for Georgian, as compared to the state-of-the-art system. In future work, we plan to explore the use of linguistic priors that apply to a group of morphologically similar low-resource languages.

Acknowledgements

This research is based upon work supported by the Intelligence Advanced Research Projects Activity (IARPA), (contract #FA8650-17-C-9117) and the National Science Foundation (awards #1941742 and #1941733). The views and conclusions herein are those of the authors and should not be interpreted as necessarily representing official policies, expressed or implied, of ODNI, IARPA, NSF or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

Ethical Considerations

The Japanese annotations were done by a linguist with appropriate compensation; we thus have ownership of the Japanese dataset for open distribution. We have been granted the rights to modify and distribute the dataset for Georgian by Eskander et al. (2020), where the annotations were done in-house by a paid linguist. Finally, the quality of the annotations was examined, both manually and empirically.

References

Christophe Andrieu, Nando De Freitas, Arnaud Doucet, and Michael I Jordan. 2003. An Introduction to MCMC for Machine Learning. *Machine learning*, 50(1-2):5–43.

Howard I Aronson. 1990. *Georgian: A Reading Grammar, Corrected Edition*. Slavica Publishers.

Jan A Botha and Phil Blunsom. 2013. Adaptor grammars for learning non- concatenative morphology. Association for Computational Linguistics.

Mathias Creutz and Krista Lagus. 2007a. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing (TSLP)*, 4(1):1–34.

Mathias Creutz and Krista Lagus. 2007b. Unsupervised models for morpheme segmentation and morphology learning. *ACM Trans. Speech Lang. Process.*, 4(1):3:1–3:34.

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.

Ramy Eskander, Owen Rambow, and Tianchun Yang. 2016. Extending the use of adaptor grammars for unsupervised morphological segmentation of unseen languages. In *Proceedings of he Twenty-Sixth International Conference on Computational Linguistics (COLING)*, Osaka, Japan.

Ramy Eskander, Owen Rambow, and Tianchun Yang. 2018. Automatically tailoring unsupervised morphological segmentation to the language. In *Proceedings of the Fifteenth SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, Brussels, Belgium.

John Goldsmith. 2001. Unsupervised learning of the morphology of a natural language. *Computational linguistics*, 27(2):153–198.

Mark Johnson, Thomas L Griffiths, and Sharon Goldwater. 2007. Adaptor grammars: A framework for specifying compositional nonparametric bayesian models. In *Advances in neural information processing systems*, pages 641–648.

Katharina Kann, Manuel Mager, Ivan Meza-Ruiz, and Hinrich Schütze. 2018. Fortification of neural morphological segmentation models for polysynthetic minimal-resource languages. *arXiv preprint arXiv:1804.06024*.

Karthik Narasimhan, Damianos Karakos, Richard M. Schwartz, Stavros Tsakalidis, and Regina Barzilay. 2014. Morphological segmentation for keyword spotting. In *EMNLP*.

Jim Pitman. 1995. Exchangeable and Partially Exchangeable Random partitions. *Probability theory and related fields*, 102(2):145–158.

Hoifung Poon, Colin Cherry, and Kristina Toutanova. 2009. Unsupervised morphological segmentation with log-linear models. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 209–217, Boulder, Colorado. Association for Computational Linguistics.

Teemu Ruokolainen, Oskar Kohonen, Sami Virpioja, and Mikko Kurimo. 2014. Painless semi-supervised morphological segmentation using conditional random fields. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 84–89.

Kairit Sirts and Sharon Goldwater. 2013. Minimally-supervised morphological segmentation using adaptor grammars. *Transactions of the Association for Computational Linguistics*, 1(May):231–242.

Sami Virpioja, Peter Smit, Stig-Arne Grönroos, Mikko Kurimo, et al. 2013. Morfessor 2.0: Python implementation and extensions for morfessor baseline.

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.