Joint Bayesian Morphology learning for Dravidian languages

Arun Kumar
Universitat Oberta de Catalunya
akallararajappan@uoc.edu

Lluís Padró
Universitat Politècnica de Catalunya
padro@cs.upc.edu

Antoni Oliver
Universitat Oberta de Catalunya
aoliverg@uoc.edu

Abstract

In this paper a methodology for learning the complex agglutinative morphology of some Indian languages using Adaptor Grammars and morphology rules is presented. Adaptor grammars are a compositional Bayesian framework for grammatical inference, where we define a morphological grammar for agglutinative languages and morphological boundaries are inferred from a plain text corpus. Once morphological segmentations are produce, regular expressions for sandhi rules and orthography are applied to achieve the final segmentation. We test our algorithm in the case of two complex languages from the Dravidian family. The same morphological model and results are evaluated comparing to other state-of-the art unsupervised morphology learning systems.

1 Introduction

Morphemes are the smallest individual units that form words. For example, the Malayalam word (മലകളട, malakaḷuṭe, related to mountains) consists of several morphemes (stem mala, plural marker kal, and genitive case marker uṭe). Morphological segmentation is one of the most studied tasks in unsupervised morphology learning (Hammarström and Borin, 2011). In unsupervised morphology learning, the words are segmented into corresponding morphemes with any supervision, as for example morphological annotations. It provides the simplest form of morphological analysis for languages that lack supervised knowledge or annotation. In agglutinative languages, there is a close connection between suffixes and morpho-syntactic functions and thus, in those languages the morphological segmentation may approximate morphological analysis well enough. Most unsupervised morphological segmentation systems have been developed and tested on a small set of European languages (Creutz and Lagus, 2007a), mainly English, Finnish and Turkish, with few exceptions in Semitic languages (Poon et al., 2009). These languages show a variety of morphological complexities, including inflection, agglutination and compounding. However, when applying those systems on other language groups with their own morphological complexities, we cannot expect the good results demonstrated so far to be automatically ported into those languages. We assume that morphological similarities of same language family enable us to define a general model that work across all languages of the family.

In this paper we work with a set of Indian languages that are highly agglutinated, with words consisting of a root and a sequence of possibly many morphemes and with each suffix corresponding to a morpho-syntactic function such as case, number, aspect, mood or gender. In addition to that, they are highly and productively compounding, allowing the formation of very long words incorporating several concepts. Thus, the morphological segmentation in those languages may partially look like a word segmentation task, which attempts to split the words in a sentence.

Dravidian languages (Steever, 2003) are a group of Indian languages that shows extensive use of morpho-phonological changes in word or morpheme boundaries during concatenation, a process called *sandhi*. This process also occurs in European languages (Andersen, 1986), but it becomes more important in the case of Dravidian languages as they use alpha-syllabic writing systems. (Taylor and Olson, 1995).

Recently, interest has shifted into semi-supervised morphological segmentation that enables to bias the model towards a certain language by using a small amount of annotated training data. We also adopt semi-supervised learning to more effectively deal with the complex orthography

of the Dravidian languages. We use the Adaptor Grammars framework (Johnson et al., 2007a) for implementing our segmentation models that has been shown to be effective for semi-supervised morphological segmentation learning (Sirts and Goldwater, 2013) and it provides a computational platform for building Bayesian non-parametric models and its inference procedure. We learn the segmentation patterns from transliterated input and convert the segmented transliterations into the orthographic representation by applying a set of regular expressions created from morphological and orthographic rules to deal with *sandhi*.

We test our system on two major languages from the Dravidian family— Malayalam and Kannada. These languages, regardless of their large number of speakers, can be considered resource-scarce, for which not much annotated data available for building morphological analyzer. We build a model that makes use of languages morphological and orthographic similarities. In Section 2, we list them main morphological and orthographic similarities of these languages.

The structure of this paper is as follows. In Section 2 we describe more thoroughly the morphological and orthographic challenges presented in Dravidian languages. Section 3 describes the Adaptor Grammars framework. In section 4, we describe the morphological segmentation system for the Dravidian languages. Experimental setup is specified in Section 5, followed by the results and analysis in section 6 and conclusions in Section 7.

2 Morphology of Dravidian languages

In this study we focus on Kannada and Malayalam, which are two major languages in the south Dravidian group. These languages are inflected and highly agglutinative, which make them morphologically complex. The writing systems of these languages are alpha-syllabic, i.e. each symbol represents a syllable. In this section we discuss morphological and orthographic similarities of these languages in detail.

2.1 Orthography

Kannada and Malayalam follow an alpha-syllabic writing system in which individual symbols are syllables. In both languages symbols are called akṣara. The atomic symbols are classified into two main categories (svaraṁ, vowels) and (vyaṅṅajnaṁ, consonants). Both languages have fourteen vowels, (including a, ā, i, ī, u, ū, e, ē, ai, o, ō, au, aṁ)[1], where aṁ is an anusvāram, which means nasalized vowel. Table 1 shows some examples of their orthographic representation. These vowels are in atomic form but when they are combined with consonant symbols, the vowel symbols change to ligatures, resulting in consonant ligatures (see examples in Table 2).

Table 1: Vowels

ISO Transliteration	a	i	o	u
Malayalam	അ	എ	ഒ	ഉ
Kannada	ಅ	ಇ	ಈ	ಉ

Table 2: Consonant ligature

	Consonant	Vowel	Ligature
ISO Transliteration	m	ī	mī
Malayalam	മ	ഈ	മീ
Kannada	ಮ	ಈ	ಮೀ

Orthography of both languages supports a large number of compound characters resulting from the combination of two consonants symbols, as those shown in Table 3.

Table 3: Compound characters

ISO Transliteration	cca	kka
Malayalam	ച്ച	ക്ക
Kannada	ಚ್ಚ	ಕ್ಕ

The orthographic systems contain characters for numerals and Arabic numerals are also present in the system. See examples in table 4

2.2 Sandhi changes

Sandhi is a morpho-phonemic change happening in the morpheme or word boundaries at the time of concatenation. Both Kannada and Malayalam have three major kinds of *sandhi* formations: deletion, insertion and assimilation. In the case of deletion *sandhi*, when two syllables are joined together one of the syllable is deleted from the resulting combination, while insertion *sandhi* adds one syllable when two syllables are joined together. The *sandhi* formations found in Sanskrit are also found in these languages as these languages loan large

[1] These symbols are according to ISO romanization standard: ISO-15919

Table 4: Numerals

Arabic Numerals	1	2
Malayalam	൧	൨
Kannada	೧	೨

number of roots from Sanskrit. There are language specific sandhi, such as visarga sandhi. Visarga sandhi is commonly found in the case of loaned words from Sanskrit. Visarga is an allophone of phonemes /r/ and /s/ at the end of the utterance. Kannada orthography keeps the symbol for visarga but Malayalam orthography uses column symbol for representing it. As languages follow alpha-syllabic orthography, these *sandhi* changes will change the orthography of the resulting word, examples are listed in Table 5. In the examples, when the Malayalam stem maḻa (mountain) joined to āṇ (is), a new syllable y introduced to resulting word maḻayāṇ. Similarly in the case of Kannada stem magu, when joined with annu, the resulting word maguyannu also includes an extra syllable y. These similarities between languages enable us to define a generic finite-state traducer for orthographic changes that are caused by phonological changes. An example of created rule for addition of syllabley is that when (a) is preceded with (u) and (a) a new syllable is y is added with resulting word.

Table 5: Sandhi Changes

Language	Insertion sandhi
Malayalam	maḻa+āṇ → maḻayāṇ
Kannada	Magu+annu → maguyannu

2.3 Morphology

Due to those orthographic properties of Dravidian languages, morpheme boundaries are marked at the syllabic level. However, during concatenation, phonological changes *sandhi* may occur, so that the resulting word has a different orthography than the individual segments concatenated together. That means the surface form may be different from the lexical form. For example, in Malayalam the surface form of the word കണ്ടില്ല (kaṇṭilla) corresponds to the lexical form കണ്ട് + ഇല്ല (kaṇṭu + illa). The changes in orthography in surface form when lexical units are combined are present in this example. Kannada also exhibits similar properties.

Malayalam and Kannada use case markers for nominative, accusative, genitive, dative, instrumental, locative, and sociative (Agesthialingom and Gowda, 1976). In both languages, nouns inflect with gender, number, and case. Gender is marked for masculine and feminine, while neuter corresponds to the absence of gender marker. There are two categories of number markers: singular and plural. Both languages can use plural markers for showing respect. For example, the Kannada word (ವಾಯು vāyu wind) is inflected with masculine gender marker. Similarly, the Malayalam word (വായു vāyu wind) has a masculine gender marker. In the case of verb morphology, both languages inflect with tenses, mood, and aspect. Where mood can follow arative, subjunctive, conditional, imperative, presumptive , abilitative, and habitual. Aspect markers can follow three categories, such as simple, progressive, purposive. For example, the Kannada sentence (ನಾನು ಬಂದು Nānu bandu I come), the verb bandu inflected with present tense marker u. Similarly the Malayalam sentence (താരം വന്നു tāraṁ vannu Star comes), the verb vannu inflected with u, which is the present tense marker.

Compounding acts as another challenge in Dravidian languages where words can have a recursive compound structure. There can be compounds embedded in a compound word, which itself can become another compound (Mohanan, 1986). For instance, in Malayalam (jātimātaviduveṣaṅṅaḷ, hatred of caste and religion) consist of first compound (jāti + māta, caste and religion), joined with other compound (viduveṣam, hatred) and plural inflection ṅṅaḷ.

3 Adaptor Grammars

Adaptor Grammars (AG) (Johnson et al., 2007a) is a non-parametric Bayesian framework for performing grammatical inference over parse trees. AG has two components—a PCFG (Probabilistic Context Free Grammar) and an adaptor function. The PCFG specifies the grammar rules used to generate the data. The adaptor function transforms the probabilities of the generated parse trees so that the probabilities of the adapted parse trees may be substantially larger than under the conditionally independent PCFG model. Various Bayesian nonparametric models can be used as Adaptor function, such as HDP (Teh et al., 2006). For instance, the Pitman-Yor Adaptor (Johnson et al., 2007a),

which is also used in this work, transforms the probability of an adapted parse tree in such a way that it is proportional to the number of times this tree has been observed elsewhere in the data. We provide an informal description of adaptor grammar here. An adaptor grammar consists of terminals V and non-terminals N, (including a start symbol, S), and initial rule set R with probability p, like a Probabilistic Context Free Grammar (PCFG). A non-terminal $A \in N$, has got a vector of concentration parameters α, where $\alpha_A > 0$. Then we say non-terminal A is adapted. If $\alpha_A = 0$ then A is an unadapted non-terminal. A non-terminal A, which is unadpted expand as in PCFG but an adapted non-terminal A can expand in two ways:

1. A can expand to a subtree t with probability $n_t/n_A + \alpha_A$, where n_t is the number of times A has expanded to t before and
2. Expand as in PCFG considering the probability propositional to concentration parameter α_A

Inference on this model can achieved using a sampling procedure . The formal definition of AGs can be found in (Johnson et al., 2007a), details of the inference procedures are described in (Johnson et al., 2007b) and (Johnson and Goldwater, 2009).

4 AGs for morphological segmentation for Dravidian languages

Dravidian languages are highly agglutinative, which means that a stem can be attached a sequence of several suffixes and several words can be concatenated together to form compounds. The segmentation model has to take these language properties into account.

We can define a grammar reflecting the agglutinative structure of language similar to the compounding grammar of (Sirts and Goldwater, 2013), excluding prefixes:

$$\begin{aligned} \text{Word} &\rightarrow \underline{\text{Compound}}^{+} \\ \underline{\text{Compound}} &\rightarrow \underline{\text{Stem}}\ \underline{\text{Suffix}}^{*} \\ \underline{\text{Stem}} &\rightarrow \underline{\text{SubMorphs}}^{+} \\ \underline{\text{Suffix}} &\rightarrow \underline{\text{SubMorphs}}^{+} \end{aligned} \tag{1}$$

Segmentation of long agglutinated sentences is the aim of above described grammar, where we consider that words are composed of words[2] and words can be composed of Stem and Suffix, where Stem and Suffix are adapted non terminals, which are "adapted" with Pitman-Yor Process (Pitman and Yor, 1997). Both Stem and Suffix can generated from drawing of Pitman-Yor process or by following PCFG rule. If these non-terminals expand according to PCFG rule, it expand to *Submorphs*, which is an intermediate levels added before terminals, For more details refer (Sirts and Goldwater, 2013)

The *Submorphs* can be defined in the following way

$$\begin{aligned} \text{Submorphs} &\rightarrow \text{Submorph} \\ \text{Submorphs} &\rightarrow \text{Submorph Submorphs} \\ \text{Submorphs} &\rightarrow \text{Chars} \\ \text{Chars} &\rightarrow \text{Char} \\ \text{Chars} &\rightarrow \text{CharChars} \end{aligned} \tag{2}$$

The Submorphs can be composed of single morph or Submorhs, which are combinations of Char. In our case Char is our internal representation for alpha-syllabic characters. The above grammar can generate various parse trees as a we put a Pitman-Yor prior on component. It is going to produce most probable morphological segmentation based on the prior probabilities. For more details of this procedure, refer (Johnson and Goldwater, 2009)

This grammar enables representing long agglutinated phrases that are common in Dravidian languages. For instance, an agglutinated Malayalam word phrase `sansthānanṅaḷileānnāṇ` with the correct morphological segmentation `sansth` + `ānanṅaḷileā` + `nnāṇ` can be represented using the grammar.

Although this grammar uses the knowledge about the agglutinative nature of the language, it is otherwise knowledge-free because it doesn't model the specific morphotactic regularities of the particular language. Next, we experiment with grammars that are specifically tailored for Dravidian languages and express the specific morphotactic patterns found in those languages. We look at the regular morphological templates described in linguistic textbooks (Krishnamurti, 1976) and (Steever, 2003) rather than generating just a sequence of generic Suffixes.

[2]The word "compound" is used in our representation for words

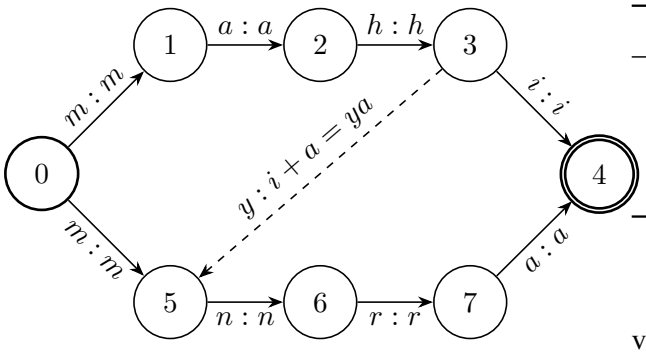

Figure 1: Example of sandhi rule in FST

4.1 Dealing with Sandhi

As explained above, the words in Dravidian languages often undergo phonological changes when morphemes or compound parts are concatenated together. Thus, in order to correctly express the segmented words in script, it is necessary to model those changes properly.

In this work we deal with *sandhi* during a post-processing step where we apply to the segmented words a set of regular expression rules that describe the valid *sandhi* rules. Our approach is similar to (Vempaty and Nagalla, 2011) where rules for orthographic changes are created. However, we use FST rules at the syllable level. Our method works with a general phonetic notation, which is the same for both languages. For example, the Malayalam word (maranna!, trees) is combination of (maram tree + nnal plural marker). We create a context sensitive rule for the orthographic change, which look like $V \rightarrow V$ m"+"||nnal. In the above V is set of all syllables in the languages. The same rule also stands for Kannada orthographic change. Similarly we create rule for orthographic changes due to *sandhi*. One example of finite-state transducer rule to handle addition sandhi is given in figure 1. It handles the `ya` sandhi happens change during the insertion sandhi. We have 62 rules for Malayalam and 34 rules for Kannada for handling sandhi changes. The statistics of the data is shown in 5.

5 Data and Experiments

We conduct our experiments on word lists extracted from Wikipedia and newspaper's websites. The statistics of the data sets are given in Table 6. Word list consist of 30 million tokens of Kannada and 40 million tokens of Malayalam. The data set consist of named entities, proper names and abbreviations. We also have 10k morphologically hand-segmented words, which act as our gold standard file.

	Kannada	Malayalam
Token frequency	30M	40M
Types	1M	1M
Labeled	10k	10k
RE Rules	62	34

Table 6: Statistics of the data sets.

In order to deal with complex orthographies of Kannada and Malayalam, we have created a internal representation, which is unique for both languages. The conversion was done in following way: the Malayalam word (അതേസമയം , atēsamayam) converted in to *a t h e s a m y a m*. During this process complex ligatures are converted into corresponding extended ASCII format and put spaces between the characters. Similarly a Kannada word (ಮಧ್ಯದ, madhyada) converted to *m a d y a d a*. This representation allow us to use the same grammar for both languages. The conversion of orthographic form to internal representation is as follows. In the first step we have converted language's scripts to corresponding ISO romanization. This representation helps in getting unique values for various ligatures and compound characters. Once the script is converted to ISO romanized form, we convert it into Extended ASCII form with unique values for each characters. As part of our experiment, we converted all words in the lists and morphological segmentations to our internal representation. For training the AG models[3], we use the scenarios proposed by (Sirts and Goldwater, 2013) we train the models using 10K, 20K 40k, 50K, 80K most frequent word types in each language with same grammar and segment the test data inductively with a parser using the AG posterior grammar as the trained model. We run five independent experiments with each setting for 1000 iterations after which we collect a single sample for getting the AG posterior grammar.

Using the trained models we segment our gold standard. Once the AG posterior grammar produce the morphological segmentation in internal form we converted internal representation into corresponding orthographic form for evaluation of result. The process of converting the internal repre-

[3]Software available at `http://web.science.mq.edu.au/~mjohnson/`

sentation to orthographic form is as follows. We take the internal representation of a word one by one and we apply finite-state rule that takes care of *sandhi* and then it convert back to orthography. The number of finite-state rules, is listed as RE rules in the Table 5.

6 Evaluation, Error analysis and Discussion

The evaluation is based on how well the methods predicts the morpheme boundaries in orthographic form and calculates precision, recall and F- score. We used Python suite provided in the morpho-challenge website[4] for evaluation purposes We also train Morfessor baseline, Morfessor-CAP and Undivide, with 80K word types. We compare our results with several baselines that have been previously successfully used for agglutinated languages: Finnish and Turkish. For unsupervised baselines we use Morfessor Categories-MAP (Creutz and Lagus, 2007b) and Undivide (Dasgupta and Ng, 2007). We train Morfessor Categories-MAP with the 80K most frequent word types and produce a model. Using this model the gold standard file is segmented and the results are compared with the manual segmentations. The same process is carried out in the case of Morfessor baseline. In the case of Undivide, we apply the system on the gold standard file and get the segmentation. We use Undivide software because it performed very well in the case of highly inflected Indian language Bengali. The results are evaluated by computing the segment boundary F1 score (F1) as is standard for morphological segmentation task.

The result achieved is presented in the table 7. In the table (P) stands for Precision and (R) stands for Recall and (F) stands for F-score.

On the manual analysis of the predicted word segmentations by our system and other baselines, we note the following:

- Our system was able to identify the sandhi changes and orthographic changes due to sandhi but other systems were unable to do that because of lack knowledge of orthography and sandhi changes.
- In the case of compound characters, Morfessor, Morfessor- MAP and Undivide segmented it into two constituent character, which is not required. For example, the Malayalam character (ങ്ക, nka) to (n) and (ka).
- All algorithms have divided compounds words.

We did not evaluated the result produced by the Adaptor Grammar individually as we need the output in language's script.

7 Conclusion and future research

We have presented a semi-supervised morphology learning technique that uses statistical measures and linguistic rules. The result of the proposed method outperforms other state-of-art unsupervised morphology learning techniques. The major contribution of this paper is the use of same model of morphology for segmenting two morphologically complex languages and the *sandhi* changes in both the languages are handled using a single finite-state transducer. In essence, we can consider it as a hybrid system, which make use of statistical information and linguistic rules together to produce better results. The experiments show that morphology of two complex languages can be learned jointly. Other important aspect of these experiments is that we tested Adaptor Grammars in the case of complex Indian languages and showed that it can be used in languages with complex morphology and orthography. The major aim of the study was to show a general model of morphology, which could be used to learn morphology of two languages. As further research, we intend to train the system with larger number of tokens and evaluate the performance in the presence of large amount of data. As we also noted an improvement in the performance when the number of word type increases.

Acknowledgments

We thank the anonymous reviewers for their careful reading of our manuscript and their many insightfulcomments and suggestions.

References

Shanmugam Agesthialingom and K Kushalappa Gowda. 1976. *Dravidian case system*. Number 39. Annamalai University.

Henning Andersen. 1986. *Sandhi phenomena in the languages of Europe*, volume 33. Walter de Gruyter.

[4] http://research.ics.aalto.fi/events/morphochallenge/

Method	Kannada			Malayalam		
	P	R	F	P	R	F
Undivide	40.98	47.17	43.86	64.08	27.12	38.22
Morfessor-base	67.92	59.02	45.63	38.21	41.59	48.54
Morfessor-MAP	70.32	53.77	62.1	62.64	47.11	53.77
Adaptor Grammar	73.63	59.82	66.01	65.66	54.32	59.45

Table 7: Results for several systems

Mathias Creutz and Krista Lagus. 2007a. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing (TSLP)*, 4(1):3.

Mathias Creutz and Krista Lagus. 2007b. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing*, 4(1):3:1–3:34.

Sajib Dasgupta and Vincent Ng. 2007. High-performance, language-independent morphological segmentation. In *HLT-NAACL*, pages 155–163.

Harald Hammarström and Lars Borin. 2011. Unsupervised learning of morphology. *Computational Linguistics*, 37(2):309–350.

Mark Johnson and Sharon Goldwater. 2009. Improving nonparametric Bayesian inference: Experiments on unsupervised word segmentation with Adaptor Grammars. In *naacl09*, pages 317–325.

Mark Johnson, Thomas L. Griffiths, and Sharon Goldwater. 2007a. Adaptor Grammars: a framework for specifying compositional nonparametric Bayesian models. In *Advances in Neural Information Processing Systems 19*, pages 641–648.

Mark Johnson, Thomas L. Griffiths, and Sharon Goldwater. 2007b. Bayesian inference for PCFGs via Markov chain Monte Carlo. In *HLT-NAACL*, pages 139–146.

Bhadriraju Krishnamurti. 1976. Comparative dravidian studies. *Current Trends in Linguistics: Index*, 14:309.

Karuvannur P Mohanan. 1986. The theory of lexical phonology: Studies in natural language and linguistic theory. *Dordrecht: D. Reidel*.

Jim Pitman and Marc Yor. 1997. The two-parameter poisson-dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, pages 855–900.

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. Association for Computational Linguistics.

Kairit Sirts and Sharon Goldwater. 2013. Minimally-supervised morphological segmentation using adaptor grammars. *Transactions of the Association for Computational Linguistics*, 1:255–266.

Sanford B Steever. 2003. *The Dravidian Languages*. Routledge.

Insup Taylor and David R Olson. 1995. *Scripts and literacy: Reading and learning to read alphabets, syllabaries, and characters*, volume 7. Springer Science & Business Media.

Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. 2006. Hierarchical dirichlet processes. *Journal of the american statistical association*, 101(476).

Phani Chaitanya Vempaty and Satish Chandra Prasad Nagalla. 2011. Automatic sandhi spliting method for telugu, an indian language. *Procedia-Social and Behavioral Sciences*, 27:218–225.