Unsupervised learning of agglutinated morphology using nested Pitman-Yor process based morpheme induction algorithm

Arun Kumar
Universitat Oberta Catalonia, UPC
arunsocs@gmail.com

Abstract

In this paper we describe a method to morphologically segment highly agglutinating and inflectional languages from Dravidian family. We use nested Pitman-Yor process to segment long agglutinated words into their basic components, and use a corpus based morpheme induction algorithm to perform morpheme segmentation. We test our method in two languages, Malayalam and Kannada and compare the results with Morfessor Categories-MAP.

1 Introduction

Morphological processing is an important task for natural language processing systems, such as information retrieval systems. In the case of languages with agglutinated and rich morphology, such as Dravidian family of languages, morphological processing is more important because one word can actually be the combination of several others, each with a number of morphological/flexive markers. Properly identifying morphemes in agglutinated words is essential for tasks such as information retrieval and machine translation.

Consider the following example from Malayalam, a language from south Dravidian family having 38 millions of native speakers and one of the classical languages of India. A word in Malayalam (പുഴകളയിരുന്നു, pul̲akaḷayirunnu, there were rivers), here root word is (പുഴ, pul̲a, river) is inflected with plural marker (കൾ, kal, Plural marker) and it also contains verb phrase(ആയിരുന്നു, ayirunnu were) all of them are joined together with orthographic changes. It is possible to have orthographic changes when words are combined, because of morpho-phonemic change called sandhi, which makes the task of segmenting Dravidian languages challenging. Dravidian languages are agglutinated like Turkish and inflected like Finnish. Other than agglutination and inflection, Orthographic changes in morpheme boundaries occurs due to sandhi changes and alpha syllabic writing system. In this case the job of a morphological analyzer is to segment the large word sequence (പുഴകളയിരുന്നു, pul̲akaḷayirunnu, there were rivers) into(പുഴ, pul̲a കൾ, kal, ആയി, ayi, ഉന്നു, unnu), which are the constituent morphemes. In the above word phrase orthography of constituent morphemes are different when they combined to in the form a word due to alpha-syllabic script that capture phonological changes. This property makes morphological processing of this languages challenging. As morpheme boundaries are marked at syllabic level, morpheme boundaries can occur inside ligatures an digraphs. In this paper we are developing a non parametric Bayesian models based on nested Pitman-Yor process on syllable level to segment long words into individual components and learn their morphological segmentation.

Dravidian family of languages are least resourced so we use corpora created from Wikipedia for conducting the experiments. We define a nested Pitman-Yor process based model for segmentation of agglutinated long sequence of words and defined model inferred using a parallel blocked Gibbs sampling algorithm. It is a generative approach in which we consider syllables are the basic units that are combined in context (agglutination) to form words. Once the algorithm achieves the segmentation on corpus created from Wikipedia, we use a heuristic search based algorithms to achieve final morphological segmentation. We test our algorithm pipeline in the case of two highly agglutinated and inflected lan-

Proceedings of the Student Research Workshop associated with RANLP 2015, pages 24–29,
Hissar, Bulgaria, 7-9 September 2015.

guages, Malayalam and Kannada from Dravidian family. As the gold standard segmentation is not available for evaluation, we created a gold standard segmentation file for both languages and evaluate the results. We manually analyze the errors in morphological segmentation to get the idea of errors that are produced by them system and to improve the system performance in further studies. In section 2 we describe previous work Bayesian non-parametric and morphological processing of agglutinating languages. In section 3 we describe Pitman-Yor models, and Section 4 describes the used algorithm for morphological segmentation. Sections 5 and 6 present the results and error analysis, and finally, section 7 presents the conclusions and future work of our research.

2 Related Work

In this section we describe related works carried out on Bayesian non-parametric models to learn morphology of languages. Research works in unsupervised learning of morphology are also relevant. Hammarström and Borin (Hammarström and Borin, 2011) provide a detailed survey of the topic. Morfessor (Creutz and Lagus, 2002; Creutz and others, 2006; Creutz et al., 2007) based on Minimum Description Length principle is the reference model for highly inflecting languages, such as Finnish. Goldwater et al. (Goldwater et al., 2009) introduce a word segmentation model based on Dirichlet Process mixture to model words and their contextual dependencies. They test their method on phonetic scripts of child speech. Following this line of research, Naradowsky & Goldwater (Naradowsky and Goldwater, 2009) incorporated English spelling rules to the morphological model to achieve better results for English phonetic script segmentation. Following these studies, Teh (Teh, 2006) introduced a Bayesian language model based on Pitman-Yor process and a new sampling procedure for the model. Lee et al. (Lee et al., 2011) modeled syntactic context to achieve better morphological segmentation. Dreyer & Eisner (Dreyer and Eisner, 2011) identified morphological paradigms using Dirichlet Process Mixture models and seed paradigms. Can and Manandhar (Can and Manandhar, 2012) clustered morphological paradigms using Hierarchical Dirichlet Process models, and Sirts & Goldwater (Sirts and Goldwater, 2013) used adapter grammar to achieve morphological segmentation. Nested Pitman-Yor process is an extension of above Dirichlet process, used to produce word segmentation of languages, such as Japanese (Mochihashi et al., 2009) and creation of language models for speech recognition (Mousa et al., 2013). These works are also relevant in the case of Bayesian non parametric models for learning morphology.

In the case of the Dravidian languages, unsupervised techniques are rarely applied. For the larger languages of the family (Telugu, Tamil, Kannada and Malayalam) there are studies that use supervised techniques. Those studies in the case of Malayalam are the following: Vasudevan & Bhattacharya (N and Bhattacharyya, 2013) propose a stemmer for Indian languages, such as Hindi, Marathi and Malayalam based on suffix lists. Idicula & David (Idicula and David, 2007) present a morphological analyzer for Malayalam based on Finite state Transducers and inflectional rules.

3 Pitman-Yor Process language model

Pitman-Yor process (Pitman, 2002) a generalization of Dirichlet process and it is a stochastic process. Goldwater et al. (Goldwater et al., 2009) and Teh (Teh, 2006) use it for language modeling. It is represented as:

$$G \sim PY(G_0, d, \theta)$$

The stochastic process generates a discrete probability distribution G similar to another given distribution $G0$. G_0 is called base measure, d is a discount factor and θ is a variable that controls similarity between both distributions G_0 and G.

A unigram language model can be expressed as a Pitman-Yor process as:

$$G_1 = p(w) \quad \forall w \in L$$

where w ranges over all words in the lexicon (L).

In the case of a bigram distribution, we have

$$G_2 = p(w|v) \quad \forall v, w \in L$$

For frequent words G_1 will be similar to G_2, so we can compute G_2 using G_1 as a base measure:

$$G_2 \sim PY(G_1, d, \theta)$$

Similarly it is possible to compute also trigram models. As this model has no analytic form the model described is represented in the form of Chinese Restaurant Process (CRP) (Aldous, 1985). Chinese Restaurant Process is an infinite large restaurant with infinitely many tables and capacity of many customers. At first the restaurant is empty, then the first customer enters and sit at an empty table. Next customer sit a new table, based on a concentration parameter or sit to already occupied table probability proportional to number of customers sitting there.

n - gram probability computed in CRP representation. Words are customers that are sitting in various tables. Tables in the restaurants are context of the words. Context of the word is length of the suffix in all earlier occurrences. So in this representation, each n-gram context h is a table and customers are n-gram counts seated over tables $1 \cdots t_{hw}$. The seat assignation to customers is constructed choosing a table k for each $c(w|h)$ (count of w given the context h) is the n- gram count and its probability is proportional to

$$p(c(w|h)) \propto \begin{cases} c_{hwk} - d, & k = (1, \cdots t_{hk}) \\ \theta + d \cdot t_h & (k = new) \end{cases}$$

where c_{hwk} is the number of customers seated in the table k and t_h is the total number of table in h. When the $k = new$, the t_h is incremented. As a result the n-gram probability can be computed as:

$$p(w|h) = \frac{c(w|h) - d \cdot t_{hw}}{\theta + c(h)} + \frac{\theta + dt_h}{\theta + c(h)} p(w|h')$$

where θ and d are the hyper parameters to be learned from data. Those parameters are inferred from the data (unsegmented corpus) and assuming that posterior probability of the variable are from Beta or Gamma distribution.

Inference on the model is done using adding and removing customers to the table t_w in the way d and θ are optimized using MCMC. For more details, refer to (Teh, 2006)

3.1 Nested Pitman-Yor process

Nested Pitman-Yor Process is a hierarchical process in which the base measure G_0 is replaced with another Pitman-Yor process. In our model base measure G_0 is replaced by a Pitman-Yor process of syllable n-grams. Then the base measure becomes:

$$G(w) = p(s_1 \cdots s_k) = \prod_{i=1}^{k} (s_i | s_{i-n+1} \cdots s_{i-1})$$

The above process can be consider as Hierarchical model, where two levels exist one is the word model and another is syllable model. We consider our syllable model as uni gram language model. For the inference it is represented in the form of a nested CRP in which a word model is connected to syllable model. In this set-up, a word w is generated from a base measure and the base measure is a Pitman - Yor process of syllables. For the inference on the particular model, we use a parallel blocked Gibbs sampler. Considering the syllables are the basic characters that joined to form words sentences. More details of sampling procedure can be found in (Neubig, 2014).

4 Morpheme identification and verification algorithm

After inference on the defined model, we apply a morpheme identification and verification algorithm to the acquired root words and morphemes. Our method is similar to that of Dasgupta & Ng (Dasgupta and Ng, 2007).

Our morpheme identification algorithm has two major parts. The first part of the algorithm is to identify a list of possible affixes for morpheme induction and composite suffixes. The list of possible affixes is extracted from the segmented corpus in following way: We assume that a word $\alpha\beta$ is concatenation of α and β, If we find both α and $\alpha\beta$ in the counter (we keep a counter of words from segmented corpus according to their frequencies) we extract β to the list of suffixes. Similarly if we find character sequence in $\alpha\beta$ and β in the counter, we list the α in the list of prefixes. But the problem with this technique is that it can create a large number of invalid suffixes and prefixes. To reduce this problem we rank the affixes based on their frequencies with different character sequences. Only top affixes that have got higher ranks are selected for induction purposes.

The second part of the algorithm aims to identify composite suffixes. As the Dravidian language family is highly inflectional large number of composite affixes are present in the vocabulary. For example in Malayalam, (ആളുകളുടെ, āḷukaḷuṭe, belongs to men) has a composite suffix (കളുടെ , kaluṭe) formed by suffixes (കൾ, kal ഉടെ, uṭe). We remove these composite suffixes from list of suffixes, otherwise it can lead to under segmentation. The third step of our morpheme identification algorithm is to identify possible roots. We take a word w from the counter and then we compose it with suffixes in the counter table. Thus, if $x + w$ (where x is an induced prefix) or $w + y$ (where x is an induced suffix) is present in the corpus, we consider w as a root and it is added to the root list. This procedure is continued until we get root, prefix and suffix lists. Using the proposed list of roots, prefixes and suffixes overall corpus is segmented to morphemes.

5 Data and Experiments

To validate our model and algorithm, we tested our algorithm on Malayalam and Kannada corpus. As Malayalam and Kannada are least resourced languages, we used a corpus crawled from Wikipedia containing 10 million words both languages, which are manually processed. As a first step of our experiments, we converted the Unicode encoded file to corresponding ISO romanized form for internal processing. We create word list of 10 million words annd add a space between characters, For example, A Kannada word (ವಿದ್ಯಾರ್ಥಿ , Vidyārthi, student) is represented as V i d y ā r t h i and it converted into constituent syllables.

Second step of the experiment consists of applying our nested Pitman-Yor model and inference algorithm to the data. For this the data is fed to the sampling algorithm for 100 iterations. Depending on the number of tokens, time taken for convergence varies. Our algorithm took 3 hours to converge in a machine with a 4-core processor with four threads in execution.

Next step is to apply our morpheme identification and evaluation algorithm to induce morpheme. Once the process is completed the system produces morphological segmentation of input words. For evaluation, we manually segmented 10,000 words of Malayalam and Kannada. The segmentation in the gold standards as follows (മനുഷ്യൻറെ,, manuṣyanṟe, of human) The segmentation is (മനുഷ്യൻ ,manuṣyan ഇൻറെ , inṟe Genitive case marker). We measured precision (P), recall (R) and F-measure (F) of predicted morpheme boundaries. We used programs provided by morpho-challenge (Virpioja et al., 2011) team for evaluation.

In order to get a comparison result, we train Morfessor Categories-MAP 0.9.2 [1] with same 10 million words for 10 Epoch and create the model. Using the model produced we segment the gold standard file and apply evaluation algorithm.

Results of the experiments shown in Table1

Table 1: Results compared to Morfessor-MAP

Method	Kannada			Malayalam		
	P	R	F	P	R	F
Morfessor- MAP	48.1	60.4	53.5	47.3	60.0	52.9
NPY	66.8	58.0	62.1	60.3	59.6	59.9

6 Error Analysis

We analyzed the results of experiments to get an insight errors that need to be solved in future research. We are listing the errors that are produced by our algorithms and Morfessor-MAP. In the case of our algorithm, it has two major steps one is to identify accurate word boundaries and other is to find accurate morpheme boundaries.

- Morfessor and our system fail to identify character combinations which need to considered as single character so it segmented digraphs and ligatures. In the case of our system it as we use a internal notation it did not segment the digraphs and ligatures.

- In the case of loaned root words, both systems fails to identify the morphemes.

[1] http://www.cis.hut.fi/projects/morpho/morfessorcatmapdownloadf

- Our system is able to identify morpheme boundaries where morpho-phonemic occurs. In the case of Morfessor-MAP, it fails to identify morpheme boundaries if there is a morpho-phonemic change and it consider zero-width joiner of Unicode as morpheme boundary.

- Our algorithms is able to identify orthographic changes that happening in the morpheme boundaries during sandhi changes but Morfessor-MAP fails. For example, a Malayalam word (മരങ്ങൾ, maraṅṅaḷ, trees) our system segment it to (മരം,maraṁ) and (ങ്ങൾ., ṅṅaḷ).

7 Conclusions and future research

We presented a method to segment words into morphemes using nested Pitman-Yor process for highly agglutinating and least resourced language such as Malayalam and Kannada. Our morphology learning system segmented complex morpheme sequences and it produce results that outperform state of the art systems. In future research, we focus on morphological processing of other languages in Dravidian family and we also focus on more richer models

Acknowledgments

References

David J Aldous. 1985. Exchangeability and related topics. In École d'Été de Probabilités de Saint-Flour XIII—1983, pages 1–198. Springer Berlin Heidelberg.

Burcu Can and Suresh Manandhar. 2012. Probabilistic hierarchical clustering of morphological paradigms. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 654–663. Association for Computational Linguistics.

Mathias Creutz and Krista Lagus. 2002. Unsupervised discovery of morphemes. In Proceedings of the ACL-02 workshop on Morphological and phonological learning-Volume 6, pages 21–30. Association for Computational Linguistics.

Mathias Creutz et al. 2006. Induction of the morphology of natural language: Unsupervised morpheme segmentation with application to automatic speech recognition. Helsinki University of Technology.

Mathias Creutz, Teemu Hirsimäki, Mikko Kurimo, Antti Puurula, Janne Pylkkönen, Vesa Siivola, Matti Varjokallio, Ebru Arisoy, Murat Saraçlar, and Andreas Stolcke. 2007. Morph-based speech recognition and modeling of out-of-vocabulary words across languages. ACM Transactions on Speech and Language Processing (TSLP), 5(1):3.

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

Markus Dreyer and Jason Eisner. 2011. Discovering morphological paradigms from plain text using a dirichlet process mixture model. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 616–627. Association for Computational Linguistics.

Sharon Goldwater, Thomas L Griffiths, and Mark Johnson. 2009. A bayesian framework for word segmentation: Exploring the effects of context. Cognition, 112(1):21–54.

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

Sumam Mary Idicula and Peter S David. 2007. A morphological processor for malayalam language. South Asia Research, 27(2):173–186.

Yoong Keok Lee, Aria Haghighi, and Regina Barzilay. 2011. Modeling syntactic context improves morphological segmentation. In Proceedings of the Fifteenth Conference on Computational Natural Language Learning, pages 1–9. Association for Computational Linguistics.

Daichi Mochihashi, Takeshi Yamada, and Naonori Ueda. 2009. Bayesian unsupervised word segmentation with nested pitman-yor language modeling. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1, pages 100–108. Association for Computational Linguistics.

Amr El-Desoky Mousa, M Ali Basha Shaik, Ralf Schlüter, and Hermann Ney. 2013. Morpheme level hierarchical pitman-yor class-based language models for lvcsr of morphologically rich languages. In INTERSPEECH, pages 3409–3413. Citeseer.

Vasudevan N and Pushpak Bhattacharyya. 2013. Little by little: Semi supervised stemming through stem set minimization. In Proceedings of the Sixth International Joint Conference on Natural Language Processing, pages 774–780, Nagoya, Japan, October. Asian Federation of Natural Language Processing.

Jason Naradowsky and Sharon Goldwater. 2009. Improving morphology induction by learning spelling rules. In IJCAI, pages 1531–1536.

Graham Neubig. 2014. Simple, correct parallelization for blocked gibbs sampling. In Technical Report, November.

Jim Pitman. 2002. Combinatorial stochastic processes. Technical report, Technical Report 621, Dept. Statistics, UC Berkeley, 2002. Lecture notes for St. Flour course.

Kairit Sirts and Sharon Goldwater. 2013. Minimally-supervised morphological segmentation using adaptor grammars. TACL, 1:255–266.

Yee Whye Teh. 2006. A hierarchical bayesian language model based on pitman-yor processes. In Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics, ACL-44, pages 985–992, Stroudsburg, PA, USA. Association for Computational Linguistics.

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.