Morphological Analysis of the Dravidian Language Family

Arun Kumar
Universitat Oberta de Catalunya
akallararajappan@uoc.edu

Ryan Cotterell
Johns Hopkins University

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

Antoni Oliver
Universitat Oberta de Catalunya
aoliverg@uoc.edu

Abstract

The Dravidian family is one of the most widely spoken set of languages in the world, yet there are very few annotated resources available to NLP researchers. To remedy this, we create DravMorph, a corpus annotated for morphological segmentation and part-of-speech. Also, we exploit novel features and higher-order models to achieve promising results on these corpora on both tasks, beating techniques proposed in the literature by as much as 4 points in segmentation F_1.

1 Introduction

The Dravidian languages comprise one of the world's major language families and are spoken by over 300 million people in southern India (see Figure 1). Despite their prevalence, they remain low resource with respect to language technology. We annotate new data and develop new models for the most commonly spoken Dravidian languages: Kannada, Malayalam, Tamil and Telugu.

We focus on the computational processing of Dravidian morphology, a critical issue since the family exhibits rich agglutinative inflectional morphology as well as highly-productive compounding. For example, Dravidian nouns are typically inflected with gender, number and case in addition to various postpositions. E.g., consider the word agniparvvatattinṟeyeāppam (അഗ്നിപർവ്വതത്തിന്റെയോപ്പം) in Malayalam which is compromised of the compound noun stem agni+paṟavvatam (*fire*+*mountain*) and the following suffixes: tta (*inflectional increment*), inṟe (*genitive case marker*), ye (*inflectional increment*) and oppam (*postposition*). These combine to give the meaning of the English phrase "with a volcano". This complexity makes morphological analysis obligatory for the Dravidian languages.

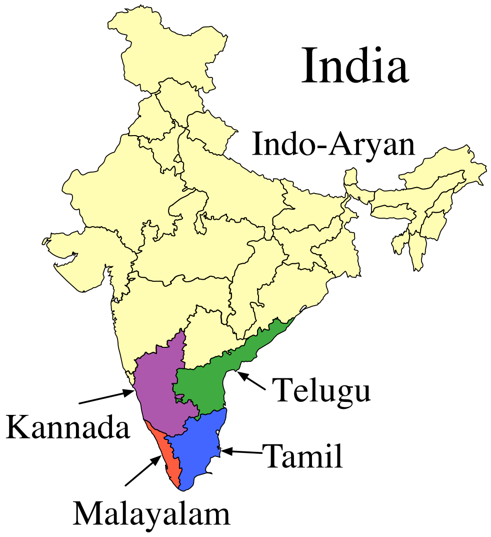

Figure 1: The Dravidian languages are spoken natively in southern India, whereas languages belonging to the Indo-Aryan family, a subbranch of the larger Indo-European family, are spoken in the north.

We make three primary contributions. (i) We release DravMorph, a corpus annotated for morphological segmentation and part-of-speech (POS) as an open-source resource, encouraging future work on Dravidian languages. (ii) We show that a combination of higher-order models and linguistically-motivated features yields state-of-the-art accuracy on the task of morphological segmentation on the corpus. (iii) We show that training POS taggers that use the output of our segmenters as features significantly improves a state-of-the-art tagger.

2 DravMorph

A primary contribution of this work is the release of DravMorph,[1] a corrected corpus for both morphological segmentation and POS in the four

[1] The morphological analyzers and the code for correcting the corpus available at https://github.com/Malkitti/Corpusandcodes

	POS	Segmentation	Wiki Dump
Ka	ILMT/IIIT-H	ILMT/IIIT-H	2015-02-09
Ma	ILMT/AM	ILMT/AM	2015-05-08
Ta	ILMT/AM	ILMT/AM	2015-05-09
Te	ILMT/AM	ILMT/UoH	2015-02-03

Table 1: The origin of the ruled-based analyzers and taggers. ILMT stands for Indian Language Machine Translation Project, AM stands for Amrita University, IIIT-H stands for IIIT-H University, UoH stands for University of Hyderabad.

	POS Tagging		Segmentation
Lang	# Sentences	# Tokens	# Types
Ka	8600	31364	3593
Ma	4034	34300	4730
Ta	4550	32400	4445
Te	5679	30625	4183

Table 2: Per language breakdown of size of the POS portion and the morphological segmentation portion of DravMorph. All train / dev / test splits used in the experiments will be released with the corpus.

most widely spoken Dravidian languages: Kannada, Malayalam, Tamil and Telugu. The corpus contains 4034-8600 annotated sentences and 3593-4730 segmented types per language. The full statistics are listed in Table 2. To the best of our knowledge, this is the most comprehensive annotated corpus of the Dravidian languages.

All the newly annotated corpora are based on Wikipedia text in the respective languages (see Table 1). To speed up annotation, we first ran closed-source ruled-based morphological analyzers and POS taggers produced by the government of India and Indian universities. We remark that the existence of such rule-based tools does not diminish the utility of the annotated corpus—our ultimate goal is the adoption of modern statistical methods for Dravidian NLP, which requires annotated data. To ensure a gold standard corpus, we then hand-corrected the resulting output. Additionally, we standardized the POS tagging schemes across languages, using the IIIT-H POS tagset (Bharati et al., 2006), which has 23 tags. Furthermore, we calculated inter-annotator agreement of two annotators for morphological labels and all datasets have Cohen's κ (Cohen, 1968) > 0.80.

3 Morphological Segmentation

We first examine the task of morphological segmentation in the Dravidian languages. The task entails breaking a word up into its constituent morphs. For example, the English word `joblessness` can be segmented as `job`+`less`+`ness`. When processing morphologically-rich languages, this helps reduce the sparsity created by the higher OOV rate due to productive morphology, and, empirically, has shown to be beneficial in a diverse variety of down-stream tasks, e.g., machine translation (Clifton and Sarkar, 2011), speech recognition (Afify et al., 2006), keyword spotting (Narasimhan et al., 2014) and parsing (Seeker and Çetinoglu, 2015). Both supervised and unsupervised approaches have been successful, but, when annotated data is available, supervised approaches typically greatly outperform unsupervised approaches (Ruokolainen et al., 2013). In light of this, we adopt a fully supervised model here.

We apply semi-Markov Conditional Random Fields (S-CRFs) to the problem of morphological segmentation (Sarawagi and Cohen, 2004; Cotterell et al., 2015). S-CRFs have the ability to jointly model both a segmentation and a labeling. For example, consider the following the Malayalam word `kūṭṭukāranmāruṭeyēāppam` (കൂട്ടുകാരന്മാരുടെയോപ്പം) (*with (male) friends*):

$$\underbrace{\textit{kūṭṭukāranmāruṭeyēāppam}}_{\boldsymbol{w}} \overset{\text{labeled segmentation}}{\Longmapsto}$$

$$\underbrace{\left[\texttt{stem}\ \textit{kūṭṭukāran}\right]}_{s_1,\ell_1} \underbrace{\left[\texttt{suf}\ \ \textit{mār}\right]}_{s_2,\ell_2} \underbrace{\left[\texttt{suf}\ \textit{uṭe}\right]}_{s_3,\ell_3} \underbrace{\left[\texttt{suf}\ \textit{yēāppam}\right]}_{s_4,\ell_4}.$$

A S-CRF models this transformation as

$$p_{\boldsymbol{\theta}}(\boldsymbol{s}, \boldsymbol{l} \,|\, \boldsymbol{w}) = \frac{1}{Z_{\boldsymbol{\theta}}(\boldsymbol{w})} \exp\left(\sum_{i=1} \boldsymbol{\theta}^\top \boldsymbol{f}(s_i, \ell_i, \ell_{i-1})\right),$$

where $\boldsymbol{s}$ is a segmentation, $\boldsymbol{\ell}$ a labeling, $\boldsymbol{\theta} \in \mathbb{R}^d$ is the parameter vector, $\boldsymbol{f}$ is a feature function[2] and the partition function $Z_{\boldsymbol{\theta}}(\boldsymbol{w})$ ensures the distribution is normalized. Note that each ℓ_i is taken from a set of labels L. In this work, we take $L = \{\texttt{prefix}, \texttt{stem}, \texttt{suffix}\}$.

As an extension to the standard S-CRF Model, we allow for higher-order segment interactions (Nguyen et al., 2011). This allows for feature functions to look at *multiple adjacent* segments s_i, s_{i-1} and s_{i-2} as well as multiple labels ℓ_i, ℓ_{i-1} and ℓ_{i-2}. While higher-order S-CRFs have shown performance improvements in various tasks, e.g., bibliography extraction and OCR (Cuong et al., 2014),

[2] Note we have omitted the dependency of $\boldsymbol{f}$ on the input $\boldsymbol{w}$ and assumed padded input for notational convenience.

they have yet to applied to morphology. We posit that the increased model expressiveness will help model more complex morphology.

We optimize the model parameters to maximize the L_2 regularized likelihood of the training data using L-BFGS (Liu and Nocedal, 1989). Computation of the likelihood and gradient can be performed efficiently through a generalization of the forward-backward algorithm that runs in $\mathcal{O}(|\boldsymbol{w}|^{n+2}|L|^{m+1})$, where n is the number of adjacent segments to be scored ($n = 0$ in a standard S-CRF) and m is the number of adjacent labels to be scored ($m = 1$ in a standard S-CRF). In this work, we explore $n \in \{0, 1, 2\}$ and $m \in \{1, 2, 3\}$, i.e., our features examine up to *three* adjacent segments and their labels.

3.1 Features

We apply a mixture of features standard for morphological segmentation and novel features based linguistic properties of the Dravidian languages.

Language Independent Feature Templates. We include the following *atomic* feature templates from Cotterell et al. (2015): (i) a binary indicator feature for substring s_i of the training data, (ii) character n-gram context features on the left and right for each potential boundary and (iii) a binary feature that fires if the segment s_i appears in a spell-checker gazetteer, to determine if it itself is a word. We also take conjunctions of all atomic features and the labels. Note that in higher-order models, we include the conjunction of all features that fire on a given segment s_i with those that fire on the adjacent segments.

Inflectional Increments. All Dravidian languages discussed in this work have semantically vacuous segments known as *inflectional increments* that are inserted during word formation between the stem and an inflectional ending. Consider the example from Malayalam, `marattinṟe` (മരത്തിന്റെ) (*tree*), which consists of stem *mara*, inflectional increment *tt* and genitive case marker *inte*. Because they *only* appear between morphs, inflectional increments serve as a cue for morph boundaries. Luckily, each set of inflectional increments is closed-class, allowing us to create a gazetteer of all increments.

Orthographic Features. The orthography of the Dravidian languages is an important factor that interacts non-trivially with the morphology. Each language uses an alpha-syllabic writing system, where each symbol encodes a *syllable*, rather than a single phoneme. Since boundaries typically occur between syllables, using a transliterated representation would throw away information. To remedy this, we include a binary feature that indicates whether a boundary corresponds to a syllable boundary in the original script. The orthographies also contain digraphs, which represent a single phoneme using a combination of two other graphs in the system. These characters are typically produced when two *syllables* are joined together at morpheme boundaries or word boundaries. Since the number of digraph characters are fixed in the orthography, we create another gazetteer for them.

Sandhi. Dravidian languages exhibit rich phonological interactions known as *sandhi* that occur at morph boundaries and word boundaries in the case of compounding. We encode the major morphophonological processes as features to capture this. We include features for the assimilation, insertion, and deletion of phonemes as these changes are visible in the surface form and can easily be represented as features. Consider an example from Malayalam, `kuṭṭiyuṁ` (കുട്ടിയും) (*child* + *also*), in this case there are two morphemes: the first morpheme `kuṭṭi`, which ends with the front vowel `i`, and the second morpheme `um`, which starts with the back vowel `u`. Sandhi inserts a glide `y` between them, marking the morpheme boundary.

4 Experiments and Results

Morphological Segmentation. On the task of morphological segmentation, we experimented with four languages from the Dravidian family in our corpus: Kannada, Malayalam, Tamil and Telugu. We first performed a full ablation study (see Table 3) on our model described in §3 to validate that both the higher-order models and the linguistic features have the desired effect. Indeed, *both* significantly improve performance. We evaluate using border F_1 (Virpioja et al., 2011) against the gold segmentation.

On test data, we compare our best system from the ablation study against two models previously proposed in the literature. First, we compare against the CRF model of Ruokolainen et al. (2013) and, second, we compare against the S-CRF model of Cotterell et al. (2015), which is just a 1st-order S-CRF. We tune the regularization coefficient for the L_2 regularizer on a held-out devel-

		Ka	*Ma*	*Ta*	*Te*
	CRF	77.09	80.44	78.02	75.88
	S-CRF $(0,1)$	77.75	80.64	78.34	76.10
2nd order	S-CRF $(1,2)$	78.49	81.05	78.75	76.64
	S-CRF $(1,2)$ +I	78.55	82.02	79.04	76.88
	S-CRF $(1,2)$ +O	78.97	82.11	79.34	76.94
	S-CRF $(1,2)$ +S	79.64	82.64	80.09	77.44
	S-CRF $(1,2)$ +I+O	79.76	82.77	80.67	77.50
	S-CRF $(1,2)$ +I+O+S	80.18	83.12	81.32	78.07
3rd order	S-CRF $(2,3)$ +I	80.34	83.26	81.40	78.77
	S-CRF $(2,3)$ +O	80.65	83. 38	81.67	78.18
	S-CRF $(2,3)$ +S	81.04	83.88	82.43	78.79
	S-CRF $(2,3)$ +I+O	82.11	84.32	82.95	78.90
	S-CRF $(2,3)$ +I+O+S	**81.24**	**85.04**	**83.90**	**79.04**

Table 3: Full ablation study on test data to test the effectiveness of our new features as well as the higher-order models. The metric used is border F_1. We denote higher-order models as S-CRF (n, m) where the integers n and m indicate the order of the model, e.g., the S-CRF $(1,2)$ models scores pairs of segments and triplets of tags. Note that +I marks *inflection increment* features, +O marks *orthography* features and +S marks *sandhi* features.

opment set.

Segmentation in POS Tagging. Next, we show the efficacy of morphological segmentation used as a preprocessing step for POS tagging (seen as a downstream task). For each type in the POS corpus, we take the MAP segmentation from the best S-CRF segmenter. We train the Marmot (Müller et al., 2013) using features derived from the segmentation. Specifically, we create a binary feature that fires on each segment in the training data. The other features in Marmot are standard shape features for POS tagging described in literature (Ratnaparkhi, 1996; Manning, 2011). Notably, the primary source of morphological information for the tagger is obtained through character n-gram features on individual word forms. Some of these features are *not* useful for the Dravidian languages, e.g., the Dravidian scripts only have lowercase.

In the Dravidian languages (and more generally agglutinative languages), morphological segments mark case, tense, aspect, gender, and number—categories indicative of the POS. For instance, tense markers only appear with verbs. These features have the potential to be *more useful* than the dynamics of the tagger as Dravidian word-order is relatively free.

Experiments and Results. We train the Marmot system with and without the morphological segmentation features. Following the procedure outlined in Müller et al. (2013), we train using stochastic gradient descent for 10 epochs with a L_1 regularizer with 0.1 coefficient. The results are reported in Table 4. We see clear gains of up to 1.69% with the systems that use the segments as features. This evinces that segmentation is a useful preprocessing step for POS tagging in Dravidian languages—character n-grams alone do not pick up on the layers of affixes.

5 Related Work

Sequence models such as CRFs and S-CRFs are used for segmentation tasks in NLP, e.g., Peng et al. (2004) applied a CRF model for Chinese word segmentation and Andrew (2006) followed with a S-CRF model. In morphology, Ruokolainen et al. (2013) train a CRF to perform morphological segmentation. Later, Ruokolainen et al. (2014) extend the work by adding semi-supervised features extracted from a large external corpus. Cotterell et al. (2015) proposed a 1st order S-CRF model for morphological segmentation, but did not explore higher-order models. Additionally, we are the first to explore rich phonological and orthographic features in supervised segmentation models.

There are large amount of research literature on construction of POS taggers for south Dravidian languages and most of them are languages specific, e.g., Pandian and Geetha (2009). However, some of the methods are applied to one or two languages in the family. PVS and Karthik (2007) apply linear-chain CRFs for POS tagging of Bengali, Hindi and Telugu. Another approach that applied to POS tagging of Dravidian language is to use part-of-speech

	Ka	*Ma*	*Ta*	*Te*
Marmot	86.35	88.77	89.04	90.50
Marmot + seg	**88.04**	**90.44**	**91.64**	**91.44**

Table 4: Tagging results using the Marmot tagger on the four Dravidian languages studied in the paper. The results indicate strongly that morphological segmentation—rather than simple prefix and suffixes n-gram features—is a useful step in handling the agglutinative Dravidian languages.

tagger of another similar languages. More recently, Kumar et al. (2015) applied adaptor grammars to unsupervised morphological segmentation of Kannada, Malayalam and Tamil.

6 Conclusion

In this paper, we presented a higher-order semi-CRF model for morphological segmentation for the Dravidian languages of South India. Our results show that the modeling of higher-order dependencies between segments and linguistically-inspired features can greatly improve system performance. We also showed that segmentation is beneficial to the down-stream task of POS tagging. To promote research on the Dravidian family, we release hand-corrected corpora for both morphological segmentation and POS tagging in four low-resource languages. Future work should concentrate on canonical segmentation (Cotterell et al., 2016a; Cotterell et al., 2016b; Cotterell and Schütze, 2017), which may be a better fit for the problem given the rich phonological changes in Dravidian morphology. Also, we plan to map the annotations to the universal POS set of Petrov et al. (2012) and the UniMorph schema of Sylak-Glassman et al. (2015).

Acknowledgments

The second author was supported by a DAAD Long-Term Research Grant and an NDSEG fellowship.

References

Mohamed Afify, Ruhi Sarikaya, Hong-Kwang Jeff Kuo, Laurent Besacier, and Yuqing Gao. 2006. On the use of morphological analysis for dialectal Arabic speech recognition. In *INTERSPEECH*.

Galen Andrew. 2006. A hybrid markov/semi-markov conditional random field for sequence segmentation. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 465–472, Sydney, Australia, July. Association for Computational Linguistics.

Akshar Bharati, Rajeev Sangal, Dipti Misra Sharma, and Lakshmi Bai. 2006. Anncorra: Annotating corpora guidelines for POS and chunk annotation for Indian languages. *LTRC-TR31*.

Ann Clifton and Anoop Sarkar. 2011. Combining morpheme-based machine translation with post-processing morpheme prediction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 32–42, Portland, Oregon, USA, June. Association for Computational Linguistics.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

Ryan Cotterell and Hinrich Schütze. 2017. Joint semantic synthesis and morphological analysis of the derived word. *CoRR*, abs/1701.00946.

Ryan Cotterell, Thomas Müller, Alexander Fraser, and Hinrich Schütze. 2015. Labeled morphological segmentation with semi-markov models. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 164–174, Beijing, China, July. Association for Computational Linguistics.

Ryan Cotterell, Arun Kumar, and Hinrich Schütze. 2016a. Morphological segmentation inside-out. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2325–2330, Austin, Texas, November. Association for Computational Linguistics.

Ryan Cotterell, Tim Vieira, and Hinrich Schütze. 2016b. A joint model of orthography and morphological segmentation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 664–669, San Diego, California, June. Association for Computational Linguistics.

Nguyen Viet Cuong, Nan Ye, Wee Sun Lee, and Hai Leong Chieu. 2014. Conditional random field with high-order dependencies for sequence labeling and segmentation. *JMLR*, 15(1):981–1009.

Arun Kumar, Lluís Padró, and Antoni Oliver. 2015. Learning agglutinative morphology of Indian languages with linguistically motivated adaptor grammars. In *RANLP*, pages 307–312, Hissar, Bulgaria.

Dong C Liu and Jorge Nocedal. 1989. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(1-3):503–528.

Christopher D Manning. 2011. Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In *CICLing*, pages 171–189. Springer.

Thomas Müller, Helmut Schmid, and Hinrich Schütze. 2013. Efficient higher-order CRFs for morphological tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 322–332, Seattle, Washington, USA, October. Association for Computational Linguistics.

Karthik Narasimhan, Damianos Karakos, Richard Schwartz, Stavros Tsakalidis, and Regina Barzilay. 2014. Morphological segmentation for keyword spotting. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 880–885, Doha, Qatar, October. Association for Computational Linguistics.

Viet Cuong Nguyen, Nan Ye, Wee Sun Lee, and Hai Leong Chieu. 2011. Semi-Markov conditional random field with high-order features. In *ICML Workshop on Structured Sparsity: Learning and Inference*.

S Lakshmana Pandian and TV Geetha. 2009. CRF models for Tamil part of speech tagging and chunking. In *International Conference on Computer Processing of Oriental Languages*, pages 11–22. Springer.

Fuchun Peng, Fangfang Feng, and Andrew McCallum. 2004. Chinese segmentation and new word detection using conditional random fields. In *Proceedings of Coling 2004*, pages 562–568, Geneva, Switzerland, Aug 23–Aug 27. COLING.

Slav Petrov, Dipanjan Das, and Ryan T. McDonald. 2012. A universal part-of-speech tagset. In *LREC*.

Avinesh PVS and G Karthik. 2007. Part-of-speech tagging and chunking using conditional random fields and transformation based learning. *Shallow Parsing for South Asian Languages*, 21.

Adwait Ratnaparkhi. 1996. A maximum entropy model for part-of-speech tagging. In *EMNLP*.

Teemu Ruokolainen, Oskar Kohonen, Sami Virpioja, and Mikko Kurimo. 2013. Supervised morphological segmentation in a low-resource learning setting using conditional random fields. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 29–37, Sofia, Bulgaria, August. 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, Gothenburg, Sweden, April. Association for Computational Linguistics.

Sunita Sarawagi and William W. Cohen. 2004. Semi-Markov conditional random fields for information extraction. In *NIPS*, pages 1185–1192.

Wolfgang Seeker and Özlem Çetinoglu. 2015. A graph-based lattice dependency parser for joint morphological segmentation and syntactic analysis. *TACL*, 3:359–373.

John Sylak-Glassman, Christo Kirov, David Yarowsky, and Roger Que. 2015. A language-independent feature schema for inflectional morphology. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 674–680, Beijing, China, July. 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. *TAL*, 52(2):45–90.