Painless Semi-Supervised Morphological Segmentation using Conditional Random Fields

Teemu Ruokolainen[a] **Oskar Kohonen**[b] **Sami Virpioja**[b] **Mikko Kurimo**[a]
[a] Department of Signal Processing and Acoustics, Aalto University
[b] Department of Information and Computer Science, Aalto University
firstname.lastname@aalto.fi

Abstract

We discuss data-driven morphological segmentation, in which word forms are segmented into morphs, that is the surface forms of morphemes. We extend a recent segmentation approach based on conditional random fields from purely supervised to semi-supervised learning by exploiting available unsupervised segmentation techniques. We integrate the unsupervised techniques into the conditional random field model via feature set augmentation. Experiments on three diverse languages show that this straightforward semi-supervised extension greatly improves the segmentation accuracy of the purely supervised CRFs in a computationally efficient manner.

1 Introduction

We discuss data-driven morphological segmentation, in which word forms are segmented into morphs, the surface forms of morphemes. This type of morphological analysis can be useful for alleviating language model sparsity inherent to morphologically rich languages (Hirsimäki et al., 2006; Creutz et al., 2007; Turunen and Kurimo, 2011; Luong et al., 2013). Particularly, we focus on a low-resource learning setting, in which only a small amount of annotated word forms are available for model training, while unannotated word forms are available in abundance.

We study morphological segmentation using conditional random fields (CRFs), a discriminative model for sequential tagging and segmentation (Lafferty et al., 2001). Recently, Ruokolainen et al. (2013) showed that the CRFs can yield competitive segmentation accuracy compared to more complex, previous state-of-the-art techniques. While CRFs yielded generally the highest accuracy compared to their reference methods (Poon et al., 2009; Kohonen et al., 2010), on the smallest considered annotated data sets of 100 word forms, they were outperformed by the semi-supervised Morfessor algorithm (Kohonen et al., 2010). However, Ruokolainen et al. (2013) trained the CRFs solely on the annotated data, without any use of the available unannotated data.

In this work, we extend the CRF-based approach to leverage unannotated data in a straightforward and computationally efficient manner via *feature set augmentation*, utilizing predictions of *unsupervised segmentation algorithms*. Experiments on three diverse languages show that the semi-supervised extension substantially improves the segmentation accuracy of the CRFs. The extension also provides higher accuracies on all the considered data set sizes and languages compared to the semi-supervised Morfessor (Kohonen et al., 2010).

In addition to feature set augmentation, there exists numerous approaches for semi-supervised CRF model estimation, exemplified by minimum entropy regularization (Jiao et al., 2006), generalized expectations criteria (Mann and McCallum, 2008), and posterior regularization (He et al., 2013). In this work, we employ the feature-based approach due to its simplicity and the availability of useful unsupervised segmentation methods. Varying feature set augmentation approaches have been successfully applied in several related tasks, such as Chinese word segmentation (Wang et al., 2011; Sun and Xu, 2011) and chunking (Turian et al., 2010).

The paper is organized as follows. In Section 2, we describe the CRF-based morphological segmentation approach following (Ruokolainen et al., 2013), and then show how to extend this approach to leverage unannotated data in an efficient manner. Our experimental setup and results are discussed in Sections 3 and 4, respectively. Finally,

we present conclusions on the work in Section 5.

2 Methods

2.1 Supervised Morphological Segmentation using CRFs

We present the morphological segmentation task as a sequential labeling problem by assigning each character to one of three classes, namely {*beginning of a multi-character morph (B)*, *middle of a multi-character morph (M)*, *single character morph (S)*}. We then perform the sequential labeling using linear-chain CRFs (Lafferty et al., 2001).

Formally, the linear-chain CRF model distribution for label sequence $y = (y_1, y_2, \ldots, y_T)$ and a word form $x = (x_1, x_2, \ldots, x_T)$ is written as a conditional probability

$$p\left(y \mid x; \boldsymbol{w}\right) \propto \prod_{t=2}^{T} \exp\Big(\boldsymbol{w} \cdot \boldsymbol{\phi}(y_{t-1}, y_t, x, t)\Big)\,, \tag{1}$$

where t indexes the character positions, $\boldsymbol{w}$ denotes the model parameter vector, and $\boldsymbol{\phi}$ the vector-valued feature extracting function. The model parameters $\boldsymbol{w}$ are estimated discriminatively based on a training set of exemplar input-output pairs (x, y) using, for example, the averaged perceptron algorithm (Collins, 2002). Subsequent to estimation, the CRF model segments test word forms using the Viterbi algorithm (Lafferty et al., 2001).

We next describe the feature set $\{\phi_i(y_{t-1}, y_t, x, t)\}_{i=1}^{|\boldsymbol{\phi}|}$ by defining *emission* and *transition* features. Denoting the label set {B, M, S} as $\mathcal{Y}$, the emission feature set is defined as

$$\{\chi_m(x,t)\mathbb{1}(y_t = y'_t) \mid m \in 1..M\,, \forall y'_t \in \mathcal{Y}\}\,, \tag{2}$$

where the indicator function $\mathbb{1}(y_t = y'_t)$ returns one if and only if $y_t = y'_t$ and zero otherwise, that is

$$\mathbb{1}(y_t = y'_t) = \begin{cases} 1 & \textbf{if} \quad y_t = y'_t \\ 0 & \textbf{otherwise} \end{cases}\,, \tag{3}$$

and $\{\chi_m(x,t)\}_{m=1}^{M}$ is the set of functions describing the character position t. Following Ruokolainen et al. (2013), we employ binary functions that describe the position t of word x using all left and right substrings up to a maximum length δ. The maximum substring length δ_{max} is considered a hyper-parameter to be adjusted using a development set. While the emission features associate the input to labels, the transition feature set

$$\{\mathbb{1}(y_{t-1} = y'_{t-1})\mathbb{1}(y_t = y'_t) \mid y'_t, y'_{t-1} \in \mathcal{Y}\} \tag{4}$$

captures the dependencies between adjacent labels as irrespective of the input x.

2.2 Leveraging Unannotated Data

In order to utilize unannotated data, we explore a straightforward approach based on feature set augmentation. We exploit *predictions of unsupervised segmentation algorithms* by defining *variants of the features* described in Section 2.1. The idea is to compensate the weaknesses of the CRF model trained on the small annotated data set using the strengths of the unsupervised methods that learn from large amounts of unannotated data.

For example, consider utilizing predictions of the unsupervised Morfessor algorithm (Creutz and Lagus, 2007) in the CRF model. In order to accomplish this, we first learn the Morfessor model from the unannotated training data, and then apply the learned model on the word forms in the annotated training set. Assuming the annotated training data includes the English word *drivers*, the Morfessor algorithm might, for instance, return a (partially correct) segmentation *driv + ers*. We present this segmentation by defining a function $\upsilon(t)$, which returns 0 or 1, if the position t is in the middle of a segment or in the beginning of a segment, respectively, as in

t	1	2	3	4	5	6	7
x_t	d	r	i	v	e	r	s
$\upsilon(t)$	1	0	0	0	1	0	0

Now, given a set of U functions $\{\upsilon_u(t)\}_{u=1}^{U}$, we define variants of the emission features in (2) as

$$\{\upsilon_u(x,t)\chi_m(x,t)\mathbb{1}(y_t = y'_t) \mid \forall u \in 1..U\,, \forall m \in 1..M\,, \forall y'_t \in \mathcal{Y}\}\,. \tag{5}$$

By adding the expanded features of form (5), the CRF model learns to associate the output of the unsupervised algorithms in relation to the surrounding substring context. Similarly, an expanded transition feature is written as

$$\{\upsilon_u(x,t)\mathbb{1}(y_{t-1} = y'_{t-1})\mathbb{1}(y_t = y'_t) \mid \forall u \in 1..U\,, \forall y'_t, y'_{t-1} \in \mathcal{Y}\}\,. \tag{6}$$

After defining the augmented feature set, the CRF model parameters can be estimated in a standard manner on the small, annotated training data set. Subsequent to CRF training, the Morfessor model is applied on the test instances in order to allow the feature set augmentation and standard decoding with the estimated CRF model. We expect the Morfessor features to specifically improve

segmentation of compound words (for example, *brain+storm*), which are modeled with high accuracy by the unsupervised Morfessor algorithm (Creutz and Lagus, 2007), but can not be learned from the small number of annotated examples available for the supervised CRF training.

As another example of a means to augment the feature set, we make use of the fact that the output of the unsupervised algorithms does not have to be binary (zeros and ones). To this end, we employ the classic letter successor variety (LSV) scores presented originally by (Harris, 1955).[1] The LSV scores utilize the insight that the predictability of successive letters should be high within morph segments, and low at the boundaries. Consequently, a high variety of letters following a prefix indicates a high probability of a boundary. We use a variant of the LSV values presented by Çöltekin (2010), in which we first normalize the scores by the average score at each position t, and subseqently logarithmize the normalized value. While LSV score tracks predictability given prefixes, the same idea can be utilized for suffixes, providing the letter predecessor variety (LPV). Subsequent to augmenting the feature set using the functions $LSV(t)$ and $LPV(t)$, the CRF model learns to associate high successor and predecessor values (low predictability) to high probability of a segment boundary. Appealingly, the Harris features can be obtained in a computationally inexpensive manner, as they merely require counting statistics from the unannotated data.

The feature set augmentation approach described above is computationally efficient, if the computational overhead from the unsupervised methods is small. This is because the CRF parameter estimation is still based on the small amount of labeled examples as described in Section 2.1, while the number of features incorporated in the CRF model (equal to the number of parameters) grows linearly in the number of exploited unsupervised algorithms.

3 Experimental Setup

3.1 Data

We perform the experiments on the Morpho Challenge 2009/2010 data set (Kurimo et al., 2009; Kurimo et al., 2010) consisting of manually prepared morphological segmentations in English, Finnish and Turkish. We follow the experiment setup, including data partitions and evaluation metrics, described by Ruokolainen et al. (2013). Table 1 shows the total number of instances available for model estimation and testing.

	English	Finnish	Turkish
Train (unann.)	384,903	2,206,719	617,298
Train (ann.)	1,000	1,000	1,000
Devel.	694	835	763
Test	10,000	10,000	10,000

Table 1: Number of word types in the Morpho Challenge data set.

3.2 CRF Feature Extraction and Training

The substring features included in the CRF model are described in Section 2.1. We include all substrings which occur in the training data. The Morfessor and Harris (successor and predecessor variety) features employed by the semi-supervised extension are described in Section 2.2. We experimented on two variants of the Morfessor algorithm, namely, the Morfessor Baseline (Creutz and Lagus, 2002) and Morfessor Categories-MAP (Creutz and Lagus, 2005), CatMAP for short. The Baseline models were trained on word types and the perplexity thresholds of the CatMAP models were set equivalently to the reference runs in Morpho Challenge 2010 (English: 450, Finnish: 250, Turkish: 100); otherwise the default parameters were used. The Harris features do not require any hyper-parameters.

The CRF model (supervised and semi-supervised) is trained using the averaged perceptron algorithm (Collins, 2002). The number of passes over the training set made by the perceptron algorithm, and the maximum length of substring features are optimized on the held-out development sets.

The experiments are run on a standard desktop computer using a Python-based single-threaded CRF implementation. For Morfessor Baseline, we use the recently published implementation by Virpioja et al. (2013). For Morfessor CatMAP, we used the Perl implementation by Creutz and Lagus (2005).

[1] We also experimented on modifying the output of the Morfessor algorithm from binary to probabilistic, but these soft cues provided no consistent advantage over the standard binary output.

3.3 Reference Methods

We compare our method's performance with the fully supervised CRF model and the semi-supervised Morfessor algorithm (Kohonen et al., 2010). For semi-supervised Morfessor, we use the Python implementation by Virpioja et al. (2013).

4 Results

Segmentation accuracies for all languages are presented in Table 2. The columns titled *Train (ann.)* and *Train (unann.)* denote the number of annotated and unannotated training instances utilized by the method, respectively. To summarize, the semi-supervised CRF extension greatly improved the segmentation accuracy of the purely supervised CRFs, and also provided higher accuracies compared to the semi-supervised Morfessor algorithm[2].

Appealingly, the semi-supervised CRF extension already provided consistent improvement over the supervised CRFs, when utilizing the computationally inexpensive Harris features. Additional gains were then obtained using the Morfessor features. On all languages, highest accuracies were obtained using a combination of Harris and CatMAP features.

Running the CRF parameter estimation (including hyper-parameters) consumed typically up to a few minutes. Computing statistics for the Harris features also took up roughly a few minutes on all languages. Learning the unsupervised Morfessor algorithm consumed 3, 47, and 20 minutes for English, Finnish, and Turkish, respectively. Meanwhile, CatMAP model estimation was considerably slower, consuming roughly 10, 50, and 7 hours for English, Finnish and Turkish, respectively. Training and decoding with semi-supervised Morfessor took 21, 111, and 47 hours for English, Finnish and Turkish, respectively.

Method	Train (ann.)	Train (unann.)	F1
English			
CRF	100	0	78.8
S-MORF.	100	384,903	83.7
CRF (Harris)	100	384,903	80.9
CRF (BL+Harris)	100	384,903	82.6
CRF (CM+Harris)	100	384,903	**84.4**
CRF	1,000	0	85.9
S-MORF.	1,000	384,903	84.3
CRF (Harris)	1,000	384,903	87.6
CRF (BL+Harris)	1,000	384,903	87.9
CRF (CM+Harris)	1,000	384,903	**88.4**
Finnish			
CRF	100	0	65.5
S-MORF.	100	2,206,719	70.4
CRF (Harris)	100	2,206,719	78.9
CRF (BL+Harris)	100	2,206,719	79.3
CRF (CM+Harris)	100	2,206,719	**82.0**
CRF	1,000	0	83.8
S-MORF.	1,000	2,206,719	76.4
CRF (Harris)	1,000	2,206,719	88.3
CRF (BL+Harris)	1,000	2,206,719	88.9
CRF (CM+Harris)	1,000	2,206,719	**89.4**
Turkish			
CRF	100	0	77.7
S-MORF.	100	617,298	78.2
CRF (Harris)	100	617,298	82.6
CRF (BL+Harris)	100	617,298	84.9
CRF (CM+Harris)	100	617,298	**85.5**
CRF	1,000	0	88.6
S-MORF.	1,000	617,298	87.0
CRF (Harris)	1,000	617,298	90.1
CRF (BL+Harris)	1,000	617,298	91.7
CRF (CM+Harris)	1,000	617,298	**91.8**

Table 2: Results on test data. *CRF (BL+Harris)* denotes semi-supervised CRF extension using Morfessor Baseline and Harris features, while *CRF (CM+Harris)* denotes CRF extension employing Morfessor CatMAP and Harris features.

5 Conclusions

We extended a recent morphological segmentation approach based on CRFs from purely supervised to semi-supervised learning. We accomplished this in an efficient manner using feature set augmentation and available unsupervised segmentation techniques. Experiments on three diverse languages showed that this straightforward semi-supervised extension greatly improves the segmentation accuracy of the supervised CRFs, while being computationally efficient. The extension also outperformed the semi-supervised Morfessor algorithm on all data set sizes and languages.

[2]The improvements over the supervised CRFs and semi-supervised Morfessor were statistically significant (confidence level 0.95) according to the standard 1-sided Wilcoxon signed-rank test performed on 10 randomly divided, non-overlapping subsets of the complete test sets.

Acknowledgements

This work was financially supported by Langnet (Finnish doctoral programme in language studies) and the Academy of Finland under the Finnish Centre of Excellence Program 2012–2017 (grant no. 251170), project *Multimodally grounded language technology* (no. 254104), and LASTU Programme (nos. 256887 and 259934).

References

Michael Collins. 2002. Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, volume 10, pages 1–8. Association for Computational Linguistics.

Çagrı Çöltekin. 2010. Improving successor variety for morphological segmentation. In *Proceedings of the 20th Meeting of Computational Linguistics in the Netherlands*.

Mathias Creutz and Krista Lagus. 2002. Unsupervised discovery of morphemes. In Mike Maxwell, editor, *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning*, pages 21–30, Philadelphia, PA, USA, July. Association for Computational Linguistics.

Mathias Creutz and Krista Lagus. 2005. Inducing the morphological lexicon of a natural language from unannotated text. In Timo Honkela, Ville Könönen, Matti Pöllä, and Olli Simula, editors, *Proceedings of AKRR'05, International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning*, pages 106–113, Espoo, Finland, June. Helsinki University of Technology, Laboratory of Computer and Information Science.

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

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*, 5(1):3:1–3:29, December.

Zellig Harris. 1955. From phoneme to morpheme. *Language*, 31(2):190–222.

Luheng He, Jennifer Gillenwater, and Ben Taskar. 2013. Graph-based posterior regularization for semi-supervised structured prediction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 38–46, Sofia, Bulgaria, August. Association for Computational Linguistics.

Teemu Hirsimäki, Mathias Creutz, Vesa Siivola, Mikko Kurimo, Sami Virpioja, and Janne Pylkkönen. 2006. Unlimited vocabulary speech recognition with morph language models applied to Finnish. *Computer Speech and Language*, 20(4):515–541, October.

Feng Jiao, Shaojun Wang, Chi-Hoon Lee, Russell Greiner, and Dale Schuurmans. 2006. Semi-supervised conditional random fields for improved sequence segmentation and labeling. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 209–216. Association for Computational Linguistics.

Oskar Kohonen, Sami Virpioja, and Krista Lagus. 2010. Semi-supervised learning of concatenative morphology. In *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, pages 78–86, Uppsala, Sweden, July. Association for Computational Linguistics.

Mikko Kurimo, Sami Virpioja, Ville Turunen, Graeme W. Blackwood, and William Byrne. 2009. Overview and results of Morpho Challenge 2009. In *Working Notes for the CLEF 2009 Workshop*, Corfu, Greece, September.

Mikko Kurimo, Sami Virpioja, and Ville Turunen. 2010. Overview and results of Morpho Challenge 2010. In *Proceedings of the Morpho Challenge 2010 Workshop*, pages 7–24, Espoo, Finland, September. Aalto University School of Science and Technology, Department of Information and Computer Science. Technical Report TKK-ICS-R37.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Carla E. Brodley and Andrea Pohoreckyj Danyluk, editors, *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289, Williamstown, MA, USA. Morgan Kaufmann.

Minh-Thang Luong, Richard Socher, and Christopher D Manning. 2013. Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning (CoNLL)*, pages 29–37. Association for Computational Linguistics, August.

Gideon Mann and Andrew McCallum. 2008. Generalized expectation criteria for semi-supervised learning of conditional random fields. In *Proceedings of ACL-08: HLT*, pages 870–878. Association for Computational Linguistics.

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.

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 (CoNLL), pages 29–37. Association for Computational Linguistics, August.

Weiwei Sun and Jia Xu. 2011. Enhancing Chinese word segmentation using unlabeled data. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 970–979. Association for Computational Linguistics.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394. Association for Computational Linguistics.

Ville Turunen and Mikko Kurimo. 2011. Speech retrieval from unsegmented Finnish audio using statistical morpheme-like units for segmentation, recognition, and retrieval. *ACM Transactions on Speech and Language Processing*, 8(1):1:1–1:25, October.

Sami Virpioja, Peter Smit, Stig-Arne Grönroos, and Mikko Kurimo. 2013. Morfessor 2.0: Python implementation and extensions for Morfessor Baseline. Report 25/2013 in Aalto University publication series SCIENCE + TECHNOLOGY, Department of Signal Processing and Acoustics, Aalto University.

Yiou Wang, Yoshimasa Tsuruoka Jun'ichi Kazama, Yoshimasa Tsuruoka, Wenliang Chen, Yujie Zhang, and Kentaro Torisawa. 2011. Improving Chinese word segmentation and POS tagging with semi-supervised methods using large auto-analyzed data. In *IJCNLP*, pages 309–317.