

# A Trie-Structured Bayesian Model for Unsupervised Morphological Segmentation

Murathan Kurfalı<sup>1</sup>, Ahmet Üstün<sup>1</sup>, and Burcu Can<sup>2</sup>

<sup>1</sup> Cognitive Science Department, Informatics Institute  
Middle East Technical University (ODTÜ)  
Ankara, 06800, Turkey

{kurfali,ustun.ahmet}@metu.edu.tr

<sup>2</sup> Department of Computer Engineering, Hacettepe University  
Beytepe, Ankara, 06800, Turkey  
burcucan@cs.hacettepe.edu.tr

**Abstract.** In this paper, we introduce a trie-structured Bayesian model for unsupervised morphological segmentation. We adopt prior information from different sources in the model. We use neural word embeddings to discover words that are morphologically derived from each other and thereby that are semantically similar. We use letter successor variety counts obtained from tries that are built by neural word embeddings. Our results show that using different information sources such as neural word embeddings and letter successor variety as prior information improves morphological segmentation in a Bayesian model. Our model outperforms other unsupervised morphological segmentation models on Turkish and gives promising results on English and German for scarce resources.

**Keywords:** unsupervised learning, morphology, morphological segmentation, Bayesian learning

## 1 Introduction

Morphological segmentation is the task of segmenting words into their meaningful units called *morphemes*. For example, the word *transformations* is split into *trans*, *form*, *ation*, and *s*. This process serves mainly as a preprocessing task in many natural language processing (NLP) applications such as information retrieval, machine translation, question answering, etc. This process is essential because sparsity becomes crucial in those NLP applications due to morphological generation that produces various word forms from a single root. It is infeasible to build a dictionary that involves all possible word forms in a language in order to use in an NLP application. Hankamer [14] suggests that the number of possible word forms in an agglutinative language such as Turkish is infinite. Therefore, instead of building a model based on word forms, morphological segmentation is applied to reduce the sparsity principally in any NLP application.

Various features have been used for morphological segmentation. Many approaches use orthographic features. However, morphology is tightly connected with syntax and semantics. Syntactic and semantic features have also been used for the segmentation task.

Features are normally used in Bayesian models in the form of a prior distribution. For example, [7] utilize frequency and length information of morphemes as prior information, which provide some orthographic features.

In this paper, we aggregate prior information from different sources in morphological segmentation within a Bayesian framework. We use orthographic features such as letter successor variety (LSV) counts obtained from tries, semantic information obtained from the neural word embeddings [17] to measure the semantic relatedness between substrings of a word, and we use the presence information of a stem in a dataset after its suffixes are stripped off assuming a concatenative morphology. Our results show that combining prior information from different sources give promising results in unsupervised morphological segmentation.

In this study, we learn tries based on semantic and orthographic features. Therefore, the output of our model is not only segmentation, but also tries that are composed of semantically and morphologically related words.

The paper is organized as follows: Section 2 presents the previous work on unsupervised morphological segmentation, section 3 defines the mathematical model, section 5 describes the inference algorithm to learn the mathematical model, section 7 presents the experimental results, and finally section 8 concludes the paper with a discussion and potential future work.

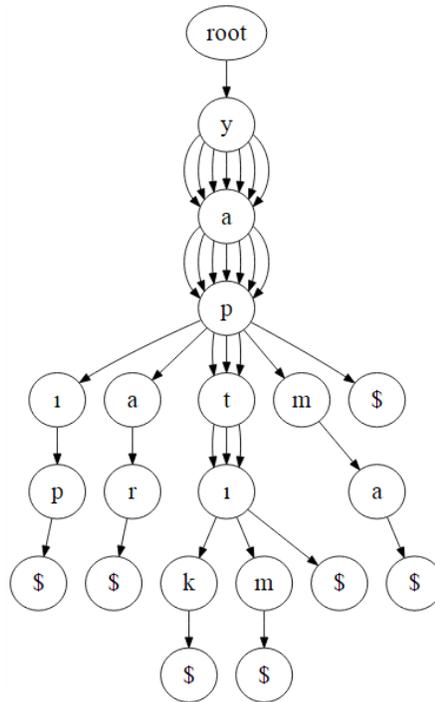
## 2 Related Work

Morphological segmentation, as one of the oldest fields in NLP, has been excessively studied. Deterministic methods are the oldest ones used in morphological segmentation. Harris [15] defines the distributional characteristics of letters in a word for the first time for unsupervised morphological segmentation. LSV model is named after Harris, which defines the morpheme boundaries based on letter successor counts. If words are inserted into a trie, branches correspond to potential morpheme boundaries. An example is given in Figure 1. In the example, *re-* is a potential prefix, and *-s*, *-ed* and *-ing* are potential suffixes in the trie due to branching that emerges before those morphemes. LSV model has been applied in various works [13,11,1,2,3]. In our study, we also use a LSV-inspired prior information, but this time in a Bayesian framework.

Stochastic methods have also been extensively used in unsupervised morphological segmentation. Morfessor is the name of the family of a group of unsupervised morphological segmentation systems which are all stochastic [8,10,9]. Non-parametric Bayesian models have also been applied in morphological segmentation [12,19,5].

Neural-inspired features are used in the recent studies. Narasimhan et al. [18] use semantic similarity obtained from neural word embeddings by word2vec [17].





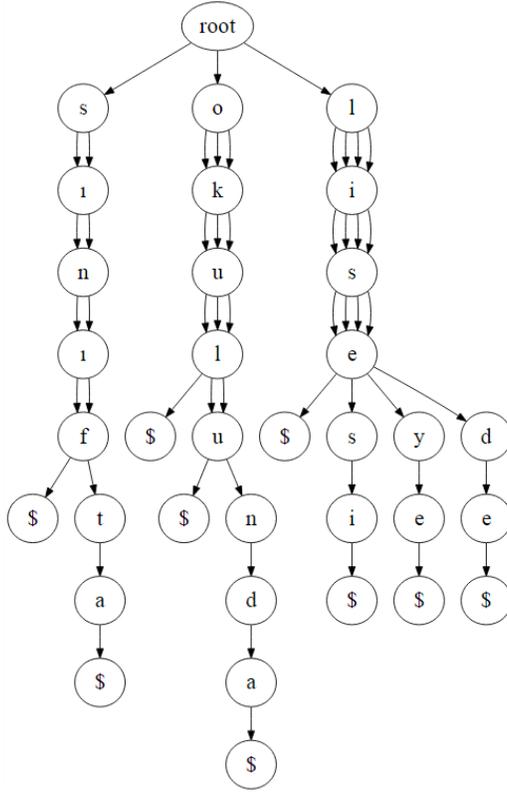
**Fig. 2.** Visualization of a trie portion that includes the word forms derived from the same stem. *yapıp*, *yapar*, *yaptık*, *yaptım*, *yapma* are inflected forms of the stem *yap* (means *to do*). The number of edges refers to the number of words in the corpus flowing in that direction on the trie. \$ denotes the end of the word.

prefix with the leftmost segmentation point in the word becomes the stem of the word.

Among the nearest 50 neighbors of the stem which are obtained from word2vec [17], the ones that begin with the same stem are inserted to the same trie. This process is repeated for each word that is inserted on the trie recursively until all the words that are semantically similar which share the same stem (detected by using the same algorithm described above) are covered. An example trie that is built with the words having the same stem is given in Figure 2.

### 3.2 Tries Based on Semantic Relatedness

Semantically related 50 words are retrieved for each word in the training set by using word2vec [17]. For each word, a trie is built and 50 similar words are inserted on the word's trie. Eventually, a trie that consists of 51 words is created for each word in the training set. A portion of a trie that involves semantically related words is given in Figure 3.



**Fig. 3.** Visualization of a trie portion built by using semantic relatedness. The trie consists of the stems *smf*, *okul*, *lise* (means *high school*, *class*, *school*) and affixed forms of these stems. The number of edges refers to the number of words in the corpus flowing in that direction on the trie. \$ denotes the end of the word.

## 4 Bayesian Model Definition

We define a Bayesian model in order to find the morpheme boundaries on the tries:

$$p(\text{Model}|\text{Corpus}) \propto p(\text{Corpus}|\text{Model})p(\text{Model}) \quad (1)$$

where *Corpus* is a list of raw words and *Model* denotes the segmentation of the corpus. The *Model* that maximizes the given posterior probability will be searched for the segmentation task. We apply a unigram model for the likelihood:

$$\begin{aligned} p(\text{Corpus}|\text{Model}) &= \prod_i^{|W|} p(w_i = (m_{i1} + m_{i2} + \dots + m_{it_i} | \text{Model})) \\ &= \prod_i^{|W|} \prod_{j=1}^{t_i} p(m_{ij} | \text{Model}) \end{aligned} \quad (2)$$

where  $w_i$  is the  $i$ th word in  $Corpus = \{w_1, \dots, w_{|W|}\}$ ,  $m_{ij}$  is the  $j$ th morpheme in  $w_i$ ,  $t_i$  is the number of morphemes in word  $w_i$ , and  $|W|$  is the number of words in the corpus. Here, morphemes are generated by a Dirichlet Process (DP) as follows:

$$m_{ij} \propto DP(\alpha, H) \quad (3)$$

with the concentration parameter  $\alpha$  and the base distribution  $H$  that is formed with a geometric distribution:

$$H(m_{ij}) = \gamma^{|m_{ij}|+1} \quad (4)$$

Here,  $|m_{ij}|$  is the length of  $m_{ij}$  and  $\gamma$  is the parameter of the geometric distribution. We assume that each letter is uniformly distributed. Therefore, we assign  $\gamma = 1/L$  where  $L$  denotes the size of the alphabet in the language. Shorter morphemes will be favored with the usage of length-inspired base distribution in the DP. From the Chinese Restaurant Process (CRP) perspective, each morpheme is generated proportionally to the number of morphemes of the same type that have already been generated (i.e. customers having the same dish):

$$p(m_{ij} = k | Model) = \frac{n_k + \alpha H(k)}{N + \alpha} \quad (5)$$

This computes the probability of  $m_{ij}$  being of type  $k$  where  $k$  refers to a distinct morpheme (i.e. morpheme type). Here,  $n_k$  is the number of morphemes of type  $k$  and  $N$  is the total number of morpheme tokens in the model. We generate each morpheme regardless of its type, such as stem, prefix, or suffix.

As for the prior information, we model the morpheme boundaries:

$$p(Model) = \prod_i^{|W|} \prod_{j=1}^{t_i} p(b_{ij}) \quad (6)$$

Here,  $b_{ij}$  refers to the  $j$ th morpheme boundary in  $w_i = m_{i1} + m_{i2} + \dots + m_{it_i}$  where  $w_i = \{b_{i1}, b_{i2}, \dots, b_{it_i}\}$ .

The probability of each  $b_{ij}$  is decomposed in terms of the number of branches leaving that node (when inserted on the trie), semantic similarity that is introduced between the two word forms that is split with  $b_{ij}$ , and the presence of the word form once the suffix is stripped off from the word:

$$p(b_{ij}) = p(b_{ij_{branch}}) \cdot p(b_{ij_{semantics}}) \cdot p(b_{ij_{presence}})$$

where  $p(b_{ij_{branch}})$  denotes the probability of  $b_{ij}$  being a morpheme boundary based on the branches leaving the trie node,  $p(b_{ij_{semantics}})$  is based on the semantic similarity of the two word forms where  $b_{ij}$  separates the two forms, and  $p(b_{ij_{presence}})$  is estimated based on the word form whether it exists in the corpus once the suffix after  $b_{ij}$  is stripped off.

Based on the LSV, the branching on the tries corresponds to the potential morpheme boundaries. We model the branching with a Poisson distribution:

$$p(b_{ij_{branch}}) = p(z_{ij} = l|\lambda) \quad (7)$$

$$= \frac{\lambda^l e^{-\lambda}}{l!} \quad (8)$$

where  $z_{ij} = l$  denotes the number of branches leaving the node below  $b_{ij}$  and  $\lambda$  is the parameter of the Poisson distribution<sup>4</sup>.

We use the cosine similarity (which is always between 0 and 1) between the neural word embeddings of the two word forms that are separated by  $b_{ij}$  for the semantic distribution:

$$p(b_{ij_{semantics}}) = \cos(x_{m_{i1}+\dots+m_{ij}}, x_{m_{i1}+\dots+m_{ij+1}}) \quad (9)$$

Here,  $x_{m_{i1}+\dots+m_{ij}}$  corresponds to the word vector of the word form  $m_{i1}+\dots+m_{ij}$  obtained from word2vec. It is the full word vector and not the compositional vector obtained from morpheme vectors.

As for the presence of the word form in the word list, we compute the likelihood of the word form  $m_{i1} + \dots + m_{ij}$ :

$$p(b_{ij_{presence}}) = \frac{f(m_{i1} + \dots + m_{ij})}{\sum_{i=1}^{|Corpus|} f(w_i)} \quad (10)$$

where  $f(m_{i1} + \dots + m_{ij})$  denotes the frequency of the word form in the corpus.

## 5 Inference

We use Gibbs sampling [6] for the inference. In each iteration, a word is uniformly selected from any trie and removed from the corpus. A binary segmentation of the word is sampled from the given posterior distribution:

$$\begin{aligned} & p(w_i = m_{i1} + m_{i2} | Corpus^{-w_i}, Model^{-w_i}, \alpha, \lambda, \gamma) \\ & \propto p(m_{i1} | Model^{-w_i}, \alpha, \gamma) p(m_{i2} | Model^{-w_i}, \alpha, \gamma) p(b_{i1}) \end{aligned} \quad (11)$$

Once a binary segmentation is sampled, another binary segmentation is sampled for  $m_{i1}$ . Therefore, a left-recursion is applied for the left part of the word. This is because of the cosine similarity that is computed between neural word embeddings of word forms and not suffixes by the original word2vec.

This process is repeated recursively until having at least 4 letters in the stem or having sampled the word itself from the posterior distribution (i.e. when the word is not segmented). An illustration is given in Figure 4.

<sup>4</sup> In the experiments, we assign  $\lambda = 4$ .

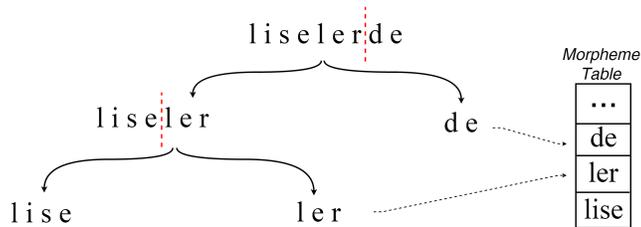


Fig. 4. The binary segmentation of the word *liselerde* (means *in the high schools*)

## 6 Segmentation

Once the model is learned, any unseen word can be segmented by using the learned model. Each word is split based on the maximum likelihood in the learned model:

$$\arg \max_{m_{i1}, \dots, m_{it_i}} p(w_i = m_{i1} + \dots + m_{it_i} | Model, \alpha, \gamma) \quad (12)$$

For the segmentation, we apply two different strategies. In both methods, we select the segmentation with the maximum likelihood, however the set of possible segmentations for the given word differs. In the first method, we only consider the segmentations learned by the model. Since the same word can exist in multiple tries, a word may have more than one different segmentation. In the second method, we consider all possible segmentations of a word and choose the one with the maximum likelihood.

## 7 Experiments and Results

We did experiments on Turkish, English and German. For each language, we built two sets of tries based on the methods described in Section 3.1 and Section 3.2. We aggregated the publicly available training and development sets provided by Morpho Challenge 2010 [16] for English, Turkish and German for training. Although gold segmentations are provided in the datasets, we only used the raw words in training. Gold segmentations were only used for evaluation purposes.

We began with 1686 English words, 1760 Turkish words, and 1779 German words obtained from the aggregated sets. Once the tries have been built by recursively augmenting the tries by using word2vec[17], eventually we obtained 2560 English word types, 43884 Turkish word types, and 13747 German word types in the tries structured from similar stems (see Section 3.1). Additionally, we obtained 34594 English word types, 67292 Turkish word types, and 23875 German types in the tries that were built based on the semantic relatedness (see Section 3.2).

We used 200-dimensional word embeddings that were obtained by training word2vec [17] on 361 million word tokens and 725.000 word types in Turkish, 129 million word tokens and 218.000 word types in English, and 651 million word

**Table 1.** Size of the datasets used in the experiments. *m1* denotes the train set built by the first method (Section 3.1) and *m2* denotes the train set built by the second method (Section 3.2)

Language	Train-m1	Train-m2	Train word2vec	Test
Turkish	43884 types	67292 types	725K types 361M tokens	1760 types
English	2560 types	34594 types	218K types 129M tokens	1686 types
German	13747 types	23875 types	608K types 651M tokens	1779 types

tokens and 608.000 word types in German. The size of all datasets used in the experiments are given in Table 1.

We compared our model with Morfessor Baseline [8] (M-Baseline), Morfessor CatMap [9] (M-CatMAP) and MorphoChain System [18]. For that purpose, we trained these models on the same training sets. We obtained the frequency information from the full word lists provided by Morpho Challenge which was need by other systems. The evaluation was performed on the aggregated training and development sets of Morpho Challenge 2010 using the Morpho Challenge evaluation method [16]. All word pairs that have a common morpheme are extracted from the results and checked whether they really share a common morpheme in the gold standard data. One point is given for each correct pair. The Precision is the proportion of the collected points to the total number of words. Recall is computed analogously. This time all word pairs that share a common morpheme are extracted from the gold standard data and checked whether they have a common morpheme in the results. For each correct pair, one point is given. Finally, the Recall is the proportion of the collected points is to the total number of words.

The results are given in Table 2 and Table 3 for tries that are composed of words structured from the same stem (see Section 3.1) and for tries that are based on semantic relatedness (see Section 3.2). According to the results, tries that contain semantically similar words achieve a better performance on morphological segmentation proving that semantically similar words also manifest similar syntactic and thus similar morphological features.

Our trie-structured model (TST) performs better than Morfessor Baseline [8], Morfessor CatMAP [9] and Morphological Chain [18] on Turkish with a F-measure of %44.16 on the tries based on semantic relatedness. We obtained a F-measure of %39.89 for Turkish from the tries structured from the same stem, which is poorer than the other method. This shows that for morphologically rich languages, semantic relatedness plays a more important role in segmentation. That is because of the sparseness of the word forms in morphologically rich languages. Here we overcome the sparsity problem with semantic information that is used in semantically built tries.

**Table 2.** Results obtained from the tries based on semantic relatedness (see Section 3.2). TST denotes our trie-structured model.

TURKISH			
	Precision (%)	Recall (%)	F-measure (%)
TST	58.27	<b>35.55</b>	<b>44.16</b>
M-CatMAP	77.78	22.91	35.40
M-Baseline	<b>84.39</b>	19.27	31.38
MorphoChain	69.45	18.29	28.95
ENGLISH			
M-Baseline	64.82	<b>64.07</b>	<b>64.44</b>
TST	56.40	47.90	51.81
MorphoChain	<b>86.26</b>	25.95	39.90
M-CatMAP	76.37	19.23	30.72
GERMAN			
M-Baseline	<b>64.74</b>	30.10	<b>41.09</b>
TST	38.66	<b>38.57</b>	38.61
M-CatMAP	62.32	15.68	25.06
MorphoChain	56.39	13.72	22.07

**Table 3.** Results obtained from the tries structured from the same stem (see Section 3.1). TST denotes our trie-structured model.

TURKISH			
	Precision (%)	Recall (%)	F-measure (%)
M-CatMAP	59.44	<b>33.41</b>	<b>42.78</b>
TST	58.85	30.17	39.89
M-Baseline	<b>74.09</b>	20.52	32.14
Morpho-Chain	72.28	25.77	38.00
ENGLISH			
M-Baseline	<b>75.28</b>	<b>61.05</b>	<b>67.42</b>
TST	58.69	51.22	54.70
MorphoChain	91.74	30.39	45.66
M-CatMAP	90.20	5.86	11.00
GERMAN			
M-Baseline	59.65	29.47	<b>39.45</b>
TST	39.62	<b>35.28</b>	37.33
MorphoChain	<b>79.06</b>	16.36	27.11
M-CatMAP	55.96	16.41	25.38

Our TST model performs better on the tries structured from the same stem on English with a F-measure of %54.70 compared to the tries based on semantic relatedness, which has a F-measure of %51.81. Since English is not a morphologically rich language, obtaining the correct stem plays an important role in segmenting the word. Words usually do not have more than one suffix and therefore finding the stem is normally sufficient to do morphological segmentation in morphologically poor languages such as English.

Our German results are close to each other obtained from two types of tries. We obtain a F-measure of %38.61 from the tries based on semantic relatedness and it performs better than Morfessor CatMAP and Morphological Chain. The F-measure is %37.33 on German, which is obtained from the tries structured from the same stem.

The results also show that Morfessor CatMAP suffers from sparsity in small datasets (especially in English), whereas our trie-structured model learns also well in small datasets.

## 8 Conclusion and Future Work

We propose a Bayesian model that utilizes semantically built trie structures that are built by using neural word embeddings (i.e. obtained from word2vec [17]) for morphological segmentation in an unsupervised setting. The current study constitutes the first part of the on-going project which in the end aims to learn part-of-speech tags and morphological segmentation jointly. To this end, the fact that the tries having semantically related words achieves the best performance paves the way of using semantically similar words in learning syntactic features.

Moreover, considering the resource-scarce languages like Turkish, our trie-structured model shows a good performance on comparably smaller datasets. In comparison to other available systems, our model outperforms them in spite of the limited training data. This shows that the small size of data can be compensated to a certain extent with structured data, that is the main contribution of this paper.

## Acknowledgments

This research is supported by the Scientific and Technological Research Council of Turkey (TUBITAK) with the project number EEEAG-115E464.

## References

1. Bordag, S.: Unsupervised knowledge-free morpheme boundary detection. In: Proceedings of the RANLP 2005 (2005)
2. Bordag, S.: Two-step approach to unsupervised morpheme segmentation. In: Proceedings of 2nd Pascal Challenges Workshop. pp. 25–29 (2006)
3. Bordag, S.: Unsupervised and knowledge-free morpheme segmentation and analysis. *Advances in Multilingual and Multimodal Information Retrieval* pp. 881–891 (2008)
4. Can, B.: Statistical Models for Unsupervised Learning of Morphology and POS tagging. Ph.D. thesis, Department of Computer Science, The University of York (2011)
5. Can, B., Manandhar, S.: Probabilistic hierarchical clustering of morphological paradigms. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. pp. 654–663. EACL '12, Association for Computational Linguistics (2012)

6. Casella, G., George, E.I.: Explaining the Gibbs sampler. *The American Statistician* 46(3), 167–174 (1992)
7. Creutz, M.: Unsupervised segmentation of words using prior distributions of morph length and frequency. In: *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*. pp. 280–287. Association for Computational Linguistics (2003)
8. Creutz, M., Lagus, K.: Unsupervised discovery of morphemes. In: *Proceedings of the ACL-02 workshop on Morphological and phonological learning*. pp. 21–30. Association for Computational Linguistics (2002)
9. Creutz, M., Lagus, K.: Inducing the morphological lexicon of a natural language from unannotated text. In: *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR 2005)*. pp. 106–113 (2005)
10. Creutz, M., Lagus, K.: Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech Language Processing* 4, 1–34 (2007)
11. Déjean, H.: Morphemes as necessary concept for structures discovery from untagged corpora. In: *Proceedings of the Joint Conferences on New Methods in Language Processing and Computational Natural Language Learning*. pp. 295–298. Association for Computational Linguistics (1998)
12. Goldwater, S., Johnson, M., Griffiths, T.L.: Interpolating between types and tokens by estimating power-law generators. In: *Advances in Neural Information Processing Systems* 18, pp. 459–466. MIT Press (2006)
13. Hafer, M.A., Weiss, S.F.: Word segmentation by letter successor varieties. *Information Storage and Retrieval* 10(11-12), 371 – 385 (1974)
14. Hankamer, J.: Finite state morphology and left to right phonology. In: *Proceedings of the West Coast Conference on Formal Linguistics*. vol. 5, pp. 41–52 (1986)
15. Harris, Z.S.: From phoneme to morpheme. *Language* 31(2), 190–222 (1955)
16. Kurimo, M., Lagus, K., Virpioja, S., Turunen, V.: Morpho Challenge 2010. <http://research.ics.tkk.fi/events/morphochallenge2010/> (2011), online; accessed 31-January-2017
17. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. *CoRR* abs/1301.3781 (2013), <http://arxiv.org/abs/1301.3781>
18. Narasimhan, K., Barzilay, R., Jaakkola, T.S.: An unsupervised method for uncovering morphological chains. *Transactions of the Association for Computational Linguistics* 3, 157–167 (2015)
19. Snyder, B., Barzilay, R.: Unsupervised multilingual learning for morphological segmentation. In: *Proceedings of ACL-08: HLT*. pp. 737–745. Association for Computational Linguistics (June 2008)
20. Soricut, R., Och, F.: Unsupervised morphology induction using word embeddings. In: *Proceedings of the Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*. pp. 1627–1637. Association for Computational Linguistics (2015)
21. Üstün, A., Can, B.: Unsupervised morphological segmentation using neural word embeddings. In: *Statistical Language and Speech Processing: 4th International Conference, SLSP 2016, Pilsen, Czech Republic, October 11-12, 2016, Proceedings*. pp. 43–53. Springer International Publishing (2016)