Semi-supervised morpheme segmentation without morphological analysis

Özkan Kılıç, Cem Bozşahin
Department of Cognitive Science
Informatics Institute, Middle East Technical University
Ankara, Turkey
E-mail: okilic@ii.metu.edu.tr, bozsahin@metu.edu.tr

Abstract

The premise of unsupervised statistical learning methods lies in a cognitively very plausible assumption that learning starts with an unlabeled dataset. Unfortunately such datasets do not offer scalable performance without some semi-supervision. We use 0.25% of METU-Turkish Corpus for manual segmentation to extract the set of morphemes (and morphs) in its 2 million word database without morphological analysis. Unsupervised segmentations suffer from problems such as oversegmentation of roots and erroneous segmentation of affixes. Our supervision phase first collects information about average root length from a small fragment of the database (5,010 words), then it suggests adjustments to structure learned without supervision, before and after a statistically approximated root, in an HMM+Viterbi unsupervised model of n-grams. The baseline of .59 f-measure goes up to .79 with just these two adjustments. Our data is publicly available, and we suggest some avenues for further research.

1. Introduction

Morpheme segmentation is the process of revealing the morphs or morphemes in a word. It can be conceived in two ways: (i) providing a sequence of morphosyntactic tags associated with the entire word, (ii) dividing the word into its morphs, with some morphic or morphemic tagging to go along with the substring covering the morph. The most common way for both tasks so far has been through morphological analysis. We describe in this work a way to approach the problem in (ii) without analysis or morphological parsing, which we tested on Turkish. As far as we know, this has been done for the first time. It uses a common Turkish language resource in two ways for the evaluation of segmenting the words into their morphemes. The first phase, semi-supervision, also yielded a gold standard of manual segmentation without labels, which we make publicly available. Although the resource is small in size (10,582 morphemes of 5,010 words), its contribution to the task is very significant, and it provides a common base for comparison in the future because the words are drawn from a well-known resource.

Rule-based morphological analyzers employ finite-state approaches with a previously compiled lexicon of morphemes. They use a set of rules for language-specific morphotactics and morpho-phonological constraints. They have been applied to concatenating languages (Koskenniemi, 1983; Hankamer, 1986; Oflazer, 1994; Çöltekin, 2010) and nonlinear templatic languages (Kiraz, 2002; Cohen-Sygal et al., 2003). Such methods are language-specific, and require their lexicons and rule sets to be updated. Statistical approaches to morpheme segmentation depend on the training of hypothetical models, which requires excessive amounts of data, from few hundred thousand to millions of words. There are well-known methods, namely supervised methods (Hajic & Hladka, 1997; Hakkani-Tür et al., 2002), unsupervised methods (Baroni et al., 2002; Creutz & Lagus, 2005; Goldwater, 2007; Yatbaz & Yuret, 2009; Yatbaz & Yuret, 2010), and semi-supervised ones (Yarowsky & Wicentowski, 2002; Kohonen et al., 2010). In these methods the training data are labeled, unlabeled, or partially labeled respectively.

Turkish is an agglutinating language with a complex morphology. Precise modeling of its morphemes using statistical methods requires large amount of data. The available resources are the METU-Turkish Corpus (Say et al., 2002) and similar academic corpora (Sak et al., 2011). This study describes an application of the Hidden Markov Model (HMM), an unsupervised method, to two million words of METU-Turkish Corpus in the first stage. The morphological tags of the corpus are ignored for unsupervised learning, and no morpheme segmentation and syntactic annotation are employed. The n-grams that form the basis for HMM are defined as states with respect to their orthographic lengths; and possible collections of orthographic representations for each n-gram are defined as emissions.

In the second stage, the model is trained on the corpus of orthographic representations of 5,010 words selected from the corpus, to calculate the initial transition and emission probabilities. These words are manually segmented by us, giving the ratio of 2 million words to 5,010 in unsupervised and supervised training. Viterbi algorithm was employed to find the most probable segmentation of a given word.

The Viterbi algorithm might suffer from the local maxima problem of the HMM. The local maxima may result from an ambiguous orthographic representation cluster which looks like a morpheme (or more precisely, a morph). It is mainly because of the tension between contrast and efficiency. Optimizing both elements gives rise to ambiguities in the collocations. For example, most of the derivational affixes and some of the inflectional affixes are frequently polysemous. For example, the suffix *-lAr* in Turkish functions as both Plu and 3P.Plu. The stem *anla* 'understand' terminates with a segment which is homographic with the inflectional suffix *-la*

(Instrumental). Similarly, the stem *ak* 'white' and the suffix *-ak* (a derivational suffix) are homographs. (Since we work on orthographic representations, we lack the phonological information such as stress to disambiguate them as homophones). As a result of such ambiguities, false segmentations such as dividing *-lar* (Plu) into *-la* (Ins) and *-r* (Aorist), and oversegmentations, e.g. dividing *kiler* 'pantry' into *-ki* (Relative) and *-ler* (Plu) do occur in the unsupervised method. Our manual segmentation of the small fragment of the database is intended to see how we can cope with these problems without attempting a morphological analysis of the test data. The method, improvements and our findings are described in the subsequent sections.

2. Method

Our HMM is a statistical model which is used to evaluate the probability of a sequence of morphemes. The model uses the Markov chain property:

• $P(s_{i,k} \mid s_{i,1}, s_{i,2}, \ldots, s_{i,k-1}) = P(s_{i,k} \mid s_{i,k-1})$

Thus the probability of next state depends only on the previous state. This seems to be a simple base to start experimenting with learning concatenative morphology. In Turkish morphotactics, the continuation of a morpheme is determined by the most recent suffix attached to a stem. For example, the suffix *-ki* can only be attached to words with either Gen or Loc, to form pronominal expressions, including inherently locative-temporal nouns such as *sabah* (morning) and *akşam* (night): *ev-de-ki* (house-Loc-ki), *ev-in-ki* (house-Gen-ki), **ev-e-ki* (house-Dat-ki), *sabah-ki* and such (Bozsahin, 2002).

In the current study, the set of states are n-grams starting from unigrams up to the longest word, and the transition probabilities are the likelihoods of possible n-gram collocations. To make the calculations easier, the 'Start' and 'End' states are inserted for each word. The emission probability of an n-gram of length-*x* is evaluated through the possible orthographic representations of length-*x* in the corpus. The Viterbi algorithm finds the optimal segmentation through the probabilities of the possible paths of the states and their emissions.

2.1 Data preparation

A subset of the METU-Sabancı Turkish Treebank (Atalay et al., 2003; Oflazer et al., 2003) is manually segmented (5,010 words). The Treebank itself consists of 7,262 annotated sentences with 43,571 words from the corpus. Both derivational and inflectional affixes are segmented. The allomorphs, such as the plural suffixes *-lar* and *-ler,* or derivational suffixes *-lik* and *-liğ,* are treated as different morphs.

In the manually segmented set, the segments with respect to their orthographic lengths correspond to n-grams in the HMM. The orthographic distributions of the n-grams lead to the emissions probabilities in the HMM. For example, the segments *-lar* and *-ın* correspond respectively to the trigram (N3) and the bigram (N2) in the HMM, and a collocation *-lar-ın* is used in estimating the transition probability from N3 to N2. In a similar manner, the orthographic representations in the manually segmented set "*-ler*, *-lar, -lik…*" and "*-in, -ın, -ün …"* are possible emissions of N3 and N2.

The statistics from the manual segmentation are used to improve the model by attempting to reduce the number of false segmentations and oversegmentations.

3. Findings

3.1 Results from the HMM

We start with the naive method of exhaustive generation of possible n-grams from the Turkish alphabet, which consists of 29 letters. No phonological filtering is applied to the n-grams before evaluating their frequencies.

The frequencies speak for themselves. For example, the most frequent n-grams in this group are inflectional morphemes, as well as some connectives and frequent function words, such as *-lar* (Plu), *ve* 'and' and *bir* 'a/one'. The least frequent n-grams are usually rare stems and nonce words, such as *ihya* 'enliven', *zzzt* and *ğaü*. Table 1 provides a summary.

	Unigram	**Bigram**	**Trigram**	**Tetragram**
Total Types	29	779	8,948	35,628
Total Tokens ~	20 million	7.5 million	6.5 million	5 million

Table 1: Total Numbers of Observed Types and Tokens of N-grams (N ≤ 4).

The most frequent 10 tokens and their percentages from about 2 million words in the corpus are given in Table 2.

Order	**Unigram**	**Bigram**	**Trigram**	**Tetragram**
1	*a*	*ar*	*lar*	*ları*
2	*e*	*la*	*ler*	*leri*
3	*n*	*an*	*eri*	*erin*
4	*r*	*er*	*arı*	*ında*
5	*i*	*le*	*bir*	*arın*
6	*l*	*in*	*ara*	*inde*
7	*k*	*de*	*nda*	*iyor*
8	*d*	*en*	*yor*	*nlar*
9	*ı*	*ın*	*ini*	*anal*
10	*m*	*da*	*ını*	*asın*
Percent in Total Tokens	53%	16.5%	6.8%	4.1%

Table 2: Most Frequent N-grams (N ≤ 4).

Figure 1 shows a very simple trellis diagram indicating the possible state transitions for the word *kedim* 'my cat', which also corresponds to possible segmentations. The emission probabilities of each n-gram and the transition probabilities among corresponding n-grams are given in Table 3 and Table 4 respectively.

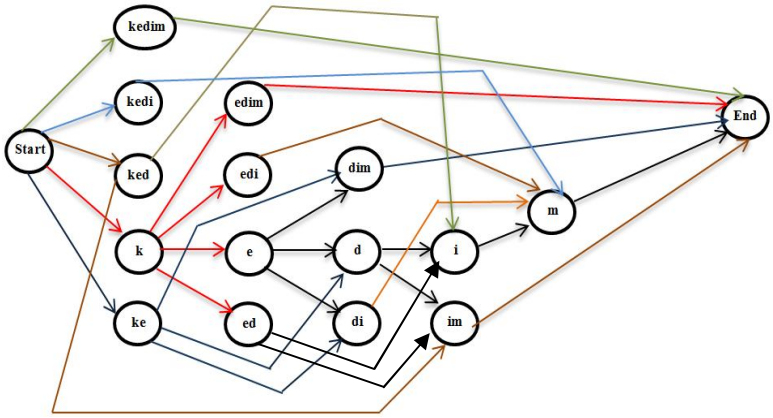

Figure 1: Trellis diagram for *kedim* 'my cat'

Output	N5	N4	N3	N2	N1	Start	End
kedim	3.81E-06						
kedi		4.73E-05					
edim		2.31E-04					
ked			1.04E-04				
edi			3.30E-03				
dim			5.43E-04				
ke				4.40E-03			
ed				4.99E-03			
di				9.22E-03			
im				4.93E-03			
k					5.23E-02		
e					7.62E-02		
d					5.08E-02		
i					7.06E-02		
m					4.12E-02		
ε						1.00E00	1.00E00

Table 3: Emission probabilities of the n-grams in the trellis diagram (Empty cells are zero. ε is empty string)

	Start	*kedim*	*kedi*	*edim*	*ked*	*edi*	*dim*	*ke*	*ed*	*di*	*im*	*k*	*e*	*d*	*i*	*m*	End
Start		1.82E-05	8.66E-03		1.87E-03			1.34E-01				4.02E-01					
kedim																	2.50E-01
kedi																1.67E-01	
edim																	2.10E-02
ked											6.02E-03				3.60E-01		
edi																5.54E-02	
dim																	2.65E-01
ke							1.19E-04			7.10E-03				1.97E-02			
ed											3.05E-02				5.51E-01		
di																4.90E-02	
im																	4.26E-01
k				1.03E-05		6.12E-04			1.70E-03				8.62E-02				
e							2.05E-03			3.70E-02				6.73E-02			
d											9.12E-03				1.86E-0		
i																7.14E-02	
m																	1.75E-01
End																	

Table 4: Transition probabilities of the n-grams in the trellis diagram (Empty cells are zero)

The Viterbi algorithm chooses the path as (Start, N4, N1, End) emitting (ε, *kedi*, *m*, ε), in which ε denotes the empty string. This is the correct sequence of morphs in the word. The second most probable path, which is slightly closer to the first path in score, is (Start, N2, N2, N1, End) emitting (ε, *ke*, *di*, *m*, ε), because of the high number of occurrences of the past tense suffix *-di* in the corpus. This is a wrong segmentation. A corpus with significantly more verbs than nouns would make the second path the winning alternative. We tried to avoid overfitting by using a representative distribution of nouns and verbs (1,414 verbs, 3,596 nouns, adjectives, adverbs and connectives). The precision, recall and f-measure values of the unsupervised method are .51, .72 and .59 respectively, which are, of course, not satisfactory.

3.2 Enhancing the Model by Manual Segmentation

The average root length of the subset we used from the Treebank is 4.09. Güngör (2003) reports the average root length to be 4.02 for Turkish. There are 150 derivational and 214 inflectional morphemes in our subset. This is the gold standard for our subset. The inflectional suffixes are very frequent. Derivational suffixes are not nearly frequent. For example, in the segmentation of the first 100 words, 59 new morphemes are discovered, of which only 6 are derivational.

To understand the cause of oversegmentations of roots by the HMM, the statistics of distinct roots whose endings are identical to morphemes in our gold standard have been evaluated from the Treebank, as shown in Table 5. For example, the most frequent root termination has the ending *-n* (10.82%).

Root Ending Segment	Percent in the Treebank
n	10.82%
k	10.13%
t	9.59%
a	9.56%
e	8.25%
r	7.69%
i	6.02%
et	4.71%
m	4.60%
an	3.86%
ş	3.18%
ı	3.04%
ol	2.41%
la	2.32%
u	2.32%
er	2.14%
le	2.00%

Table 5: Percentages of some root endings with morpheme-like segments

We incorporate this edge statistic to our experiments as follows: if the sum of the indices of visited states (a measure of length) is close to the calculated average root length 4.09, and if in the current state a symbol identical to one of our morpheme endings *x* from Table 5 is observed, then the state's transition probability is multiplied by (1- percentage-of-*x*), which gives the probability of *x* not being an edge of the roots from the Treebank. For example, if a unigram is in the 4th orthographic position of a word and it emits *-n*, then its transition probability is multiplied by (1 - 0.1082). This is a simple way to check the effect of the edge statistic on oversegmentation of roots, because it forces the Viterbi algorithm to favor likely endings of roots and morphemes.

Next we tackle the false segmentation problem of morphemes. The statistics from the segmented subset are used for this purpose to look at structure past the average root length. For example, *-lArI* (3Plu.Poss) and *-lar-I* (Plu-Acc) are identical orthographically, hence they are prone to false segmentation. Manual segmentations show that there are 190 occurrences of the latter one, of which 59% have at least one more segment before the word boundary. On the other hand, 3Plu.Poss occurs in 40 words of which 30% are in word boundaries.

The statistics of such problematic cases were part of our experiments. Their (1- 'edge probabilities') are multiplied with the transition probabilities of the HMM considering the locations and emission types of the states. For example, if *-lArI* has the transition probability .085, and *-lAr* .075, and if 70% of *-lArI* are not at the word boundary compared to 59% for *-lAr*, determined from supervision, the numbers (1-.7)x.085 and (1-.59)x.075 would be the contenders. By doing so, the Viterbi algorithm is partially directed to a path starting with a 3-gram (Plu) instead of a 4-gram (3Plu.Poss) for *-lArI-* representations occurring before the word boundaries.

4. Results and Conclusion

The working principles in our two experiments are to disfavor oversegmentations of roots and false segmentations of affixes by incorporating the collocations of root endings and morpheme starts. Employing this much semi-supervision from a very small fragment (0.25%) of the database successfully increased the measures to (.72, .87, .79) (precision, recall, f-measure), from (.51, .72, .59) of the unsupervised method, over 2 million unlabeled words. Considering the knowledge-poor strategies we employed, and the fact that we did nothing to reveal the structure in compounds, this is quite striking, and shows us more avenues to move toward unsupervised segmentation. (1,838 words, out of 5,010 are either roots or compounds, which seems to be a representative percentage). We also note that we get .79 f-measure of correct segmentation into morphemes, i.e. we deliver the morphs, without morphological analysis, not just the overall tag for the word. What manual segmentation provides is syntactic and semantic disambiguation in an indirect way, hence some semantic-phonological cues (such as intonation, stress) and some limited syntactic knowledge (e.g. for compounds), are next targets we want to address. Our manual segmentation is going to be made publicly available at www.LcsL.metu.edu.tr/share. It will in time

grow up to a size of 45,000 words. We plan to take on MorphoChallenge data after this level of supervision.

5. References

Atalay, N. B., Oflazer, K., Say, B. (2003). The Annotation Process in the Turkish Treebank. In *Proceedings of the EACL Workshop on Linguistically Interpreted Corpora - LINC*, April 13-14, 2003, Budapest, Hungary.

Baroni, M., Matiasek, J., Trost, H. (2002). Unsupervised discovery of morphologically related words based on orthographic and semantic similarity. *In Proceedings of the Workshop on Morphological and Phonological Learning at ACL '02*, pp. 11–20.

Bozsahin, C. (2002). The Combinatory Morphemic Lexicon. *Computational Linguistics*, *28*(2), pp. 145-186.

Cohen-Sygal, Y., Gerdemann, D., Wintner, S. (2003). Computational Implementation of Non-Concatenative Morphology. In *Proceedings of the Workshop on Finite-State Methods in Natural Language Processing, an EACL'03 Workshop*, pp. 59-66, Budapest, Hungary.

Çöltekin, Ç. (2010). A Freely Available Morphological Analyzer for Turkish. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC2010)*, Valletta, Malta, May 2010.

Creutz, M., Lagus, K. (2005). 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*, Espoo, pp. 106-113.

Goldwater, S. J. (2007). Nonparametric Bayesian models of lexical acquisition. Ph.D. thesis. Brown University, Providence, RI, USA.

Güngör, T. (2003). Lexical and Morphological Statistics for Turkish. In *Proceedings of TAINN* 2003, pp. 409-412.

Hajic, J., Hladka, B. (1997). Tagging of inflective languages: a comparison. In *Proceedings of ANLP'97*, Washington, DC, pp. 136–143. ACL.

Hakkani-Tür, D. Z., Oflazer, K., Tür, G. (2002). Statistical morphological disambiguation for agglutinative languages. *Computers and the Humanities, 36*(4), pp. 381–410.

Hankamer, J. (1986). Finite state morphology and left to right phonology. In *Proceedings of the West Coast Conference on Formal Linguistics*, Volume 5. Stanford Linguistic Association.

Kiraz, G. A. (2002). *Computational Nonlinear Morphology, with Emphasis on Semitic Languages.* Cambridge, U.K.: Cambridge University Press.

Kohonen, O., Virpioja, S., Lagus, K. (2010). Semi-supervised learning of concatenative morphology. In *Proceedings of the Eleventh Meeting of the ACL* Special Interest Group on Computational Phonology and Morphology (SIGMORPHON 2010), Uppsala, pp. 78–86.

Koskenniemi, K. (1983). Two-level morphology: A general computational model for word-form recognition and generation. Ph.D. Thesis. University of Helsinki.

Oflazer, K. (1994). Two-level description of Turkish morphology. *Literary and Linguistic Computing*, 9(2).

Oflazer, K., Say, S., Hakkani-Tür, D. Z., Tür, G. (2003) Building a Turkish Treebank, Invited chapter in A. Abeille (Ed.), *Building and Exploiting Syntactically-annotated Corpora*. Kluwer Academic Publishers.

Sak, H., Güngör, T., Saraçlar, M. (2011). Resources for Turkish Morphological Processing. *Language Resources and Evaluation, 45*(2), pp. 249–261.

Say, B., Zeyrek, D., Oflazer, K., Özge, U. (2002). Development of a corpus and a treebank for present-day written Turkish. In *Proceedings of the Eleventh International Conference of Turkish Linguistics*.

Yarowsky, D., Wicentowski, R. (2000). Minimally supervised morphological analysis by multimodal alignment. In *Proceedings of the ACL*, pp. 207–216.

Yatbaz, M. A., Yuret, D. (2009). Unsupervised Morphological Disambiguation using Statistical Language Models. In *Proceedings of the NIPS 2009 Workshop on Grammar Induction, Representation of Language and Language Learning.*

Yatbaz, M. A., Yuret, D. (2010). Unsupervised Part of Speech Tagging Using Unambiguous Substitutes from a Statistical Language Model. In *Proceedings of the Coling 2010*, Beijing, pp. 1391–1398.