Exploiting Morphology in Turkish Named Entity Recognition System

Reyyan Yeniterzi*
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA, 15213, USA
reyyan@cs.cmu.edu

Abstract
Turkish is an agglutinative language with complex morphological structures, therefore using only word forms is not enough for many computational tasks. In this paper, we analyze the effect of morphology in a Named Entity Recognition system for Turkish. We start with the standard word-level representation and incrementally explore the effect of capturing syntactic and contextual properties of tokens. Furthermore, we also explore a new representation in which roots and morphological features are represented as separate tokens instead of representing only words as tokens. Using syntactic and contextual properties with the new representation provides an 7.6% relative improvement over the baseline.

1 Introduction
One of the main tasks of information extraction is the Named Entity Recognition (NER) which aims to locate and classify the named entities of an unstructured text. State-of-the-art NER systems have been produced for several languages, but despite all these recent improvements, developing a NER system for Turkish is still a challenging task due to the structure of the language.

Turkish is a morphologically complex language with very productive inflectonal and derivational processes. Many local and non-local syntactic structures are represented as morphemes which end produces Turkish words with complex morphological structures. For instance, the following English phrase “if we are going to be able to make [something] acquire flavor” which contains the necessary function words to represent the meaning can be translated into Turkish with only one token “tatlandırabileceksek” which is produced from the root “tat” (flavor) with additional morphemes +lan (acquire), +dr (to make), +abil (to be able), +cek (are going), +se (if) and +k (we).

This productive nature of the Turkish results in production of thousands of words from a given root, which causes data sparseness problems in model training. In order to prevent this behavior in our NER system, we propose several features which capture the meaning and syntactic properties of the token in addition to the contextual properties. We also propose using a sequence of morphemes representation which uses roots and morphological features as tokens instead of words.

The rest of this paper is organized as follows: Section 2 summarizes some previous related works, Section 3 describes our approach, Section 4 details the datasets used in the paper, Section 5 reports the experiments and results and Section 6 concludes with possible future work.

2 Related Work
The first paper (Cucerzan and Yarowski, 1999) on Turkish NER describes a language independent bootstrapping algorithm that learns from word internal and contextual information of entities. Turkish was one of the five languages the authors experimented with. In another work (Tur et al., 2003),
the authors followed a statistical approach (HMMs) for NER task together with some other Information Extraction related tasks. In order to deal with the agglutinative structure of the Turkish, the authors worked with the root-morpheme level of the word instead of the surface form. A recent work (Küçük and Yazici, 2009) presents the first rule-based NER system for Turkish. The authors used several information sources such as dictionaries, list of well known entities and context patterns.

Our work is different from these previous works in terms of the approach. In this paper, we present the first CRF-based NER system for Turkish. Furthermore, all these systems used word-level tokenization but in this paper we present a new tokenization method which represents each root and morphological feature as separate tokens.

3 Approach

In this work, we used two tokenization methods. Initially we started with the sequence of words representation which will be referred as word-level model. We also introduced morpheme-level model in which morphological features are represented as states. We used several features which were created from deep and shallow analysis of the words. During our experiments we used Conditional Random Fields (CRF) which provides advantages over HMMs and enables the use of any number of features.

3.1 Word-Level Model

Word-level tokenization is very commonly used in NER systems. In this model, each word is represented with one state. Since CRF can use any number of features to infer the hidden state, we develop several feature sets which allow us to represent more about the word.

3.1.1 Lexical Model

In this model, only the word tokens are used in their surface form. This model is effective for many languages which do not have complex morphological structures. However for morphologically rich languages, further analysis of words is required in order to prevent data sparseness problems and produce more accurate NER systems.

3.1.2 Root Feature

An analysis (Hakkani-Tür, 2000) on English and Turkish news articles with around 10 million words showed that on the average 5 different Turkish word forms are produced from the same root. In order to decrease this high variation of words we use the root forms of the words as an additional feature.

3.1.3 Part-of-Speech and Proper-Noun Features

Named entities are mostly noun phrases, such as first name and last name or organization name and the type of organization. This property has been used widely in NER systems as a hint to determine the possible named entities.

Part-of-Speech tags of the words depend highly on the language and the available Part-of-Speech tagger. Taggers may distinguish the proper nouns with or without their types. We used a Turkish morphological analyzer (Of lazer, 1994) which analyzes words into roots and morphological features. An example to the output of the analyzer is given in Table 1. The part-of-speech tag of each word is also reported by the tool 1. We use these tags as additional features and call them part-of-speech (POS) features.

The morphological analyzer has a proper name database, which is used to tag Turkish person, location and organization names as proper nouns. An example name entity with this +Prop tag is given in Table 1. Although, the use of this tag is limited to the given database and not all named entities are tagged with it, we use it as a feature to distinguish named entities. This feature is referred as proper-noun (Prop) feature.

3.1.4 Case Feature

As the last feature, we use the orthographic case information of the words. The initial letter of most named entities is in upper case, which makes case feature a very common feature in NER tasks. We also use this feature and mark each token as UC or LC depending on the initial letter of it. We don’t do

---

1 The meanings of various Part-of-Speech tags are as follows: +A3pl - 3rd person plural; +P3sg - 3rd person singular possessive; +Gen - Genitive case; +Prop - Proper Noun; +A3sg - 3rd person singular; +Pnon - No possessive agreement; +Nom - Nominative case.
anything special for the first words in sentences.

An example phrase in word-level model is given in Table 2. In the figure each row represents a state. The first column is the lexical form of the word and the rest of the columns are the features and the tag is in the last column.

3.2 Morpheme-Level Model

Using Part-of-Speech tags as features introduces some syntactic properties of the word to the model, but still there is missing information of other morphological tags such as number/person agreements, possessive agreements or cases. In order to see the effect of these morphological tags in NER, we propose a morpheme-level tokenization method which represents a word in several states; one state for a root and one state for each morphological feature.

In a setting like this, the model has to be restricted from assigning different labels to different parts of the word. In order to do this, we use an additional feature called root-morph feature. The root-morph is a feature which is assigned the value “root” for states containing a root and the value “morph” for states containing a morpheme. Since there are no prefixes in Turkish, a model trained with this feature will give zero probability (or close to zero probability if there is any smoothing) for assigning any B-* (Begin any NE) tag to a morph state. Similarly, transition from a state with B-* or I-* (Inside any NE) tag to a morph state with O (Other) tag will get zero probability from the model.

In morpheme-level model, we use the following features:

- the actual root of the word for root and morphemes of the token
- the Part-of-speech tag of the word for the root part and the morphological tag for the morphemes
- the root-morph feature which assigns “root” to the roots and “morph” to the morphemes
- the proper-noun feature
- the case feature

An example phrase in root-morpheme-based chunking is given in Table 3. In the figure each row represents a state and each word is represented with several states. The first row of each word contains the root, POS tag and Root value for the root-morph feature. The rest of the rows of the same word contains the morphemes and Morph value for the root-morph feature.

4 Data Set

We used training set of the newspaper articles data set that has been used in (Tur et al., 2003). Since we do not have the test set they have used in their paper, we had to come up with our own test set. We used only 90% of the train data for training and left the remaining for testing.

Three types of named entities; person, organization and location, were tagged in this dataset. If the word is not a proper name, then it is tagged with other. The number of words and named entities for each NE type from train and tests sets are given in Table 4.

Table 4: The number of words and named entities in train and test set

|        | #WORDS | #PER. | #ORG. | #LOC. |
|--------|--------|-------|-------|-------|
| **TRAIN** | 445,498 | 21,701 | 14,510 | 12,138 |
| **TEST**  | 47,344 | 2,400  | 1,595  | 1,402  |

5 Experiments and Results

Before using our data in the experiments we applied the Turkish morphological analyzer tool (Oflazer, 1994) and then used Morphological disambiguator (Sak et al., 2008) in order to choose the correct morphological analysis of the word depending on the

---

1. Table 1: Examples to the output of the Turkish morphological analyzer

| WORD + ROOT + POS + MORPHEMES |
|-------------------------------|
| beyinlerinin (of their brains) + beyin + Noun + A3pl+P3sg+Gen |
| Amerika (America) + Amerika + Noun + Prop+A3sg+Pnon+Nom |

2. One can see that Ilias which is Person NE is not tagged as Prop (Proper Noun) in the example, mainly because it is missing in the proper noun database of the morphological analyzer.
context. In experiments, we used CRF++ ³, which is an open source CRF sequence labeling toolkit and we used the conlleval ⁴ evaluation script to report F-measure, precision and recall values.

5.1 Word-level Model

In order to see the effects of the features individually, we inserted them to the model one by one iteratively and applied the model to the test set. The F-measures of these models are given in Table 5. We can observe that each feature is improving the performance of the system. Overall the F-measure was increased by 6 points when all the features are used.

5.2 Morpheme-level Model

In order to make a fair comparison between the word-level and morpheme-level models, we used all the features in both models. The results of these experiments are given in Table 6. According to the table, morpheme-level model achieved better results than word-level model in person and location entities. Even though word-level model got better F-Measure score in organization entity, morpheme-level is much better than word-level model in terms of recall.

Using morpheme-level tokenization to introduce morphological information to the model did not hurt the system, but it also did not produce a significant improvement. There may be several reasons for this. One can be that morphological information is not helpful in NER tasks. Morphemes in Turkish words are giving the necessary syntactic meaning to the word which may not be useful in named entity finding. Another reason for not seeing a significant change with morpheme usage can be our representation. Dividing the word into root and morphemes and using them as separate tokens may not be the best way of using morphemes in the model. Other ways of representing morphemes in the model may produce more effective results.

As mentioned in Section 4, we do not have the same test set that has been used in Tur et al. (Tur et al., 2003). Even though it is impossible to make a fair comparison between these two systems, it would

---

³CRF++: Yet Another CRF toolkit
⁴www.cnts.ua.ac.be/conll2000/chunking/conlleval.txt
Table 5: F-measure Results of Word-level Model

|                  | PERSON | ORGANIZATION | LOCATION | OVERALL |
|------------------|--------|--------------|----------|---------|
| LEXICAL MODEL (LM) | 80.88  | 77.05        | 88.40    | 82.60   |
| LM + ROOT         | 83.32  | 80.00        | 90.30    | 84.96   |
| LM + ROOT + POS   | 84.91  | 81.63        | 90.18    | 85.98   |
| LM + ROOT + POS + PROP | 86.82  | 82.66        | 90.52    | 87.18   |
| LM + ROOT + POS + PROP + CASE | 88.58  | 84.71        | 91.47    | 88.71   |

Table 6: Results of Morpheme-Level (Morp) and Word-Level Models (Word)

|                  | PRECISION | RECALL | F-MEASURE |
|------------------|-----------|--------|-----------|
|                  | MORP      | WORD   | MORP      | WORD   |
| PERSON           | 91.87%    | 91.41% | 86.92%    | 85.92% |
| ORGANIZATION     | 85.23%    | 91.00% | 81.84%    | 79.23% |
| LOCATION         | 94.15%    | 92.83% | 90.23%    | 90.14% |
| OVERALL          | 91.12%    | 91.81% | 86.87%    | 85.81% |

Table 7: F-measure Comparison of two systems

|                  | OURS  | (TUR ET AL., 2003) |
|------------------|-------|-------------------|
| BASELINE MODEL   | 82.60 | 86.01             |
| BEST MODEL       | 88.94 | 91.56             |
| IMPROVEMENT      | 7.6%  | 6.4%              |

be good to note how these systems performed with respect to their baselines which is lexical model in both. As it can be seen from Table 7, both models improved upon their baselines significantly.

6 Conclusion and Future Work

In this paper, we explored the effects of using features like root, POS tag, proper noun and case to the performance of NER task. All these features seem to improve the system significantly. We also explored a new way of including morphological information of words to the system by using several tokens for a word. This method produced compatible results to the regular word-level tokenization but did not produce a significant improvement.

As future work we are going to explore other ways of representing morphemes in the model. Here we represented morphemes as separate states, but including them as features together with the root state may produce better models. Another approach we will also focus is dividing words into characters and applying character-level models (Klein et al., 2003).

Acknowledgments

The author would like to thank William W. Cohen, Kemal Of lazer, Gökhan Tur and Behrang Mohit for their valuable feedback and helpful discussions. The author also thank Kemal Of lazer for providing the data set and the morphological analyzer. This publication was made possible by the generous support of the iLab and the Center for the Future of Work. The statements made herein are solely the responsibility of the author.

References

Silviu Cucerzan and David Yarowski. 1999. Language independent named entity recognition combining morphological and contextual evidence. In Proceedings of the Joint SIGDAT Conference on EMNLP and VLC, pages 90–99.

Dilek Z. Hakkani-Tür. 2000. Statistical Language Modelling for Turkish. Ph.D. thesis, Department of Computer Engineering, Bilkent University.

Dan Klein, Joseph Smarr, Huy Nguyen, and Christopher D. Manning. 2003. Named entity recognition with character-level models. In Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4, pages 180–183.

Dilek Küçük and Adnan Yazici. 2009. Named entity recognition experiments on Turkish texts. In Proceedings of the 8th International Conference on Flexible Query Answering Systems, FQAS ’09, pages 524–535, Berlin, Heidelberg. Springer-Verlag.

Kemal Of lazer. 1994. Two-level description of Turk-
ish morphology. *Literary and Linguistic Computing*, 9(2):137–148.

Hasım Sak, Tunga Güngör, and Murat Saracılı. 2008. Turkish language resources: Morphological parser, morphological disambiguator and web corpus. In *Advances in Natural Language Processing*, volume 5221 of *Lecture Notes in Computer Science*, pages 417–427.

Gökhan Tur, Dilek Z. Hakkani-Tür, and Kemal Of lazer. 2003. A statistical information extraction system for Turkish. In *Natural Language Engineering*, pages 181–210.