Morphological Embeddings for Named Entity Recognition in Morphologically Rich Languages

Onur Güngör
Boğaziçi University, Istanbul
Huawei R&D Center, Istanbul
onurgu@boun.edu.tr

Eray Yıldız
Huawei R&D Center, Istanbul
eray.yildiz@huawei.com

Suzan Üsküdarlı
Boğaziçi University, Istanbul
suzan.uskudarli@boun.edu.tr

Tunga Güngör
Boğaziçi University, Istanbul
gungort@boun.edu.tr

Abstract

In this work, we present new state-of-the-art results of 93.59% and 79.59% for Turkish and Czech named entity recognition based on the model of (Lample et al., 2016). We contribute by proposing several schemes for representing the morphological analysis of a word in the context of named entity recognition. We show that a concatenation of this representation with the word and character embeddings improves the performance. The effect of these representation schemes on the tagging performance is also investigated.

1 Introduction

Named Entity Recognition (NER) is an important task in Natural Language Processing (NLP) that aims to discover references to entities in some text. Identified entities are classified into predefined categories like person, location and organization. NER is mostly utilized prior to complex natural language understanding tasks such as relation extraction, knowledge base generation, and question answering (Liu and Ren, 2011; Lee et al., 2007). Additionally, NER systems are often part of search engines (Guo et al., 2009b) and machine translation systems (Babych and Hartley, 2003).

Early studies regarding NER propose using hand crafted rules and lists of names of people, places and organizations (Humphreys et al., 1998; Appelt et al., 1995). Traditional approaches typically use several hand-crafted features such as capitalization, word length, gazetteer related features, and syntactic features (part-of-speech tags, chunk tags, etc.) (McCallum and Li, 2003; Finkel et al., 2005). A wide range of machine learning-based methods have been proposed to address named entity recognition. Some of the well known approaches are conditional random fields (CRF) (McCallum and Li, 2003; Finkel et al., 2005), maximum entropy (Borthwick, 1999), bootstrapping (Jiang and Zhai, 2007; Wu et al., 2009), latent semantic association (Guo et al., 2009a), and decision trees (Szarvas et al., 2006).

Recently, deep learning models have been instrumental in deciding how the parts of the input should be composed to allow the most beneficial features to form leading to state-of-the-art results (Collobert et al., 2011). Likewise, researchers have found that representing words with fixed length vectors in a dense space helps improving the overall performance of many tasks: sentiment analysis (Socher et al., 2013), syntactic parsing (Collobert and Weston, 2008), language modeling (Mikolov et al., 2010), part-of-speech tagging and NER (Collobert et al., 2011). These word representations or embeddings are automatically learned both during or before the training using various methods such as Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014).

Building upon these findings, there are recent studies which treat the NER task as a sequence labeling problem which employ LSTM or GRU components (Lample et al., 2016; Huang et al., 2015; Ma and Hovy, 2016; Yang et al., 2016) to capture the syntactic and semantic relations between the units that make up a natural language sentence. However, these approaches are not well studied for morphologically rich languages. Unlike other languages, morphologically rich languages such as Turkish may retain important information in the morphology of the surface form of the word while the same information may be contained in the syntax of other languages. For example, the word "İstanbul'daydı" means 'he/she was in Istanbul' in English. The morphological analysis of the word

is "İstanbul+Noun+Prop+A3sg+Pnon+LocˆDB+Verb+Zero+Past+A3sg" where 'Prop' indicates a proper noun, 'A3sg' signifies the third singular person aggreement whereas 'Pnon' signifies no possesive agreement is active. 'DB' indicates a transition of Part-Of-Speech type usually induced by a derivative suffix (Oflazer, 1994). In this case, the derivation is triggered by the '-di' suffix which was decoded as 'Past' tag which indicates past tense. As seen from the example, morphological tags may help in capturing syntactic and semantic information. In order to address this, character based embeddings in word representations (Lample et al., 2016) and entities tagged at character level (Kuru et al., 2016) were proposed for NER. Embedding based frameworks for representing morphology were also proposed in other contexts such as language modeling (Luong et al., 2013; dos Santos and Zadrozny, 2014; Xu and Liu, 2017; Bhatia et al., 2016; Lankinen et al., 2016) and morphological tagging and segmentation (Shen et al., 2016; Cotterell and Schütze, 2017). However, even though morphological tags have been employed in the past (Tür et al., 2003; Yeniterzi, 2011), our work is the first to propose an embedding based framework for representing the morphological analysis in the context of NER.

We build upon a state-of-the-art NER tagger (Lample et al., 2016) based on a sequential neural model with extensible word representations in Section 2. We show that augmenting the word representation with morphological embeddings based on Bi-LSTMs (Section 2.1) improves the performance of the base model, which uses only pretrained word embeddings. We contribute by investigating various configurations of the morphological analysis of the surface form of a word (Section 2.2). In Section 3.3, we compare the performance of several experiment setups which employ character and morphological embeddings in various combinations. We report F1-measures of 93.59% and 79.59% for Turkish and Czech respectively. These results are the state-of-the-art results compared to previous work (Demir and Özgür, 2014; Seker and Eryiğit, 2012) which rely on a regularized averaged perceptron and CRF respectively both with hand crafted features.

2 Model

We formally define an input sentence as $X = (x_1, x_2, \ldots, x_n)$ where each x_i is a fixed length vector of size d, consisting of embeddings that represent the ith word (See Section 2.1). x_i are then fed to a Bi-LSTM which is composed of two LSTMs (Hochreiter and Schmidhuber, 1997) treating the input forwards and backwards respectively. Thus we obtain these forward and backward components' cell matrices $\overrightarrow{H}$ and $\overleftarrow{H}$ which are both of size $n \times p$, where p is the number of dimensions of one component of the Bi-LSTM. Thus, $\overrightarrow{H}_{i,j}$ is the value of jth dimension of ith output vector of the right component which corresponds to the ith word in the sentence. We then feed the concatenation of these matrices $H = [\overrightarrow{H}, \overleftarrow{H}]$ to a fully connected hidden layer of K output neurons.

To model the dependencies between the corresponding labels of consecutive input units, we follow a conditional random field (CRF) (Lafferty et al., 2001) based approach. This dependency is clearly indicated by labels in IOB tagging scheme, i.e. B-PERSON, I-PERSON, etc.

To do this, we obtain a score vector at each position i and aim to minimize the following objective function for a single sample sentence X:

$$s(X, y) = \sum_i A_{y_i, y_{i+1}} + \sum_i \xi_{i, y_i}$$

where $A_{i,j}$ represents the score of a transition from tag i to j and ξ_i are the tag scores at position i output by the uppermost fully connected layer. Using this model, we decode the most probable tagging sequence y^* as $\arg\max_{\tilde{y}} s(X, \tilde{y})$.

2.1 Embeddings

It has been shown that modeling units of information in a natural language input as fixed length vectors is more effective at encoding semantic properties of the words compared to deciding on the features apriori (Collobert et al., 2011; Turian et al., 2010). Therefore we represent the input words, x_i, as a combination of three embeddings: *word*, *character*, and *morphological*. Thus the size d of x_i is $d_w + 2d_m + 2d_c$. We describe these embeddings below and illustrate in Figure 1.

Word embeddings. A vector of size d_w which is learned by the global objective function. However, we never learn this component from scratch, instead we load pretrained vectors.

Character embeddings. We learn another fixed length vector of size $2d_c$ for each word. However, in contrast with a word embedding, we

want to capture the covert relationships in the sequence of characters of the word. To achieve this, we have a separate Bi-LSTM component for this embedding type with a cell dimension of d_c. We feed it with the characters of the surface form of x_i and concatenate the cell output of the forward and backward LSTMs to obtain the *character* embedding of the word.

Morphological embeddings. These are constructed similar to *character* embeddings. In this case, the individual tags of the morphological analysis are treated as a sequence and fed into the separate Bi-LSTM component for *morphological* embeddings to obtain a vector of length $2d_m$. We devised several different combinations of morphological tags which is explained in Section 2.2. To illustrate, we use the word 'evlerinde' which can both mean 'in their house' or 'in their houses' or 'in his/her houses' in Turkish. In Figure 1, we assume that the correct morphological analysis is 'ev+Noun+A3pl+P3sg+Loc', where 'A3pl' indicates 3rd person plural, 'P3sg' is the possessive marker for 3rd person singular, and 'Loc' is the locative case marker, thus can be translated as 'in his/her houses'.

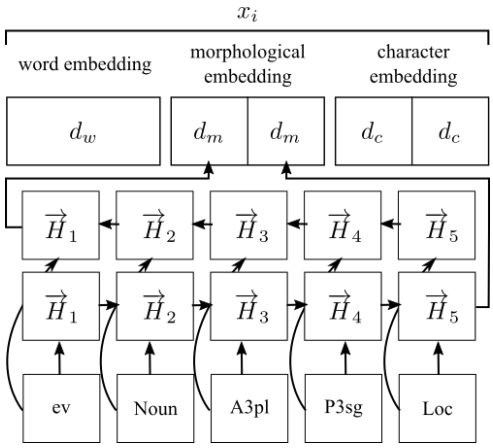

Figure 1: Bi-LSTM component for morphological embeddings.

2.2 Embedding configurations

We experimented with different combinations of morphological tags to discover an effective configuration for extracting the syntactic and semantic information in the morphological analysis of a token.

A simple embedding configuration is to use all morphological tags in the analysis along with or without the root (we call this WR and WOR respectively). Secondly we tried to remove the tags between the root and the derivation boundary (DB) based on the information they carry may not be relevant in some aspects because of the transformation into a new word with partly different lexical and syntactic properties. We call this version WR_ADB, i.e. "İstanbul+^DB+Verb+Zero+Past+A3sg". Lastly, we devised a scheme in which we treat the string of morphological analysis as a surface form and process each character of this surface form as we did for *character* embeddings and call this scheme as CHAR.

3 Experiments

3.1 Training

The parameters to be learned by the training algorithm is the parameters of the Bi-LSTM in Section 2, the parameters of the Bi-LSTMs for the *character* and *morphological* embeddings and the word embeddings for each unique word. We experimented with several different choices for the number of dimensions for these parameters and observed that a choice of 100 for word embeddings, 200 for character embeddings and 200 for morphological embeddings. We trained the models by calculating the gradients using backpropagation algorithm and updating with stochastic gradient descent algorithm with a learning rate of 0.01. We also employed gradient clipping to handle gradients diverging from zero. We additionally used dropout on the inputs with probability 0.5.

3.2 Dataset

We train and evaluate our model with a corpus which is widely used in previous work on Turkish NER (Tür et al., 2003). In addition to the entity tags and tokens, the corpus also contains the disambiguated morphological tags of input words. We observed many morphological analysis errors and incorrect entity taggings in the corpus (Tür et al., 2003), probably as a result of automated analysis and labeling.

We obtained word embeddings[1] of Turkish words as vectors of length 100 using the skipgram algorithm (Mikolov et al., 2013) on a corpus of 951M words (Yildiz et al., 2016), 2,045,040 of which are unique. This corpus consists of Turkish text extracted from several national newspapers, news sites, and book transcripts.

[1]We will make these word embeddings available at (we refrain from sharing the url during the review process).

Turkish

	WE (pretrained)		
Setup No	**CE**	**ME**	**F1**
1	-	-	90.96
2	-	ME(WR_ADB)	90.76
3	-	ME(WR)	90.33
4	-	ME(CHAR)	92.79
5	-	ME(WOR)	91.18
6	CE	ME(WR_ADB)	91.37
7	CE	ME(WR)	91.09
8	CE	ME(CHAR)	92.93
9	CE	ME(WOR)	**93.59**
10	CE	-	93.37
Czech			
11	CE	ME(CHAR)	**79.59**

Table 1: Best performing experiment setups.

3.3 Results

We observed that using pretrained word embeddings gave the best results compared to learning word embeddings while training the model. Therefore we only include the results of experiment setups with pretrained word embeddings in Table 1.

We start with comparing Setup 1 and 4 (and 5) and suggest that the morphological analysis does indeed contribute to higher performance with CHAR and WOR. However, Setup 2 and 3 did not reach the performance level of Setup 1. We suspect that the reason is the relatively high number of parameters when we include the 20030 roots into the model. This effect is also seen in Setup 6 and 7 which have a lower performance compared to Setups 8 and 9. However, we have to note that using character embeddings alone also improved the performance in Setup 10. Nevertheless, we see that the best performance is achieved in Setup 9 when both of them are employed. However, when we performed the McNemar's test (Dietterich, 1998), we observed that the difference between them is not significant at 95% confidence level. We explain the difference in performance between Setup 4 and 5 with the errors in the morphological analysis which are mostly due to unknown or mispelled words. In those cases, the analysis become usually the same nominal case with 3rd person singular. We suspect that the fact that ME(CHAR) can process the root even if it is faulty allows it to capture useful information into the embedding.

Despite CE caused a large improvement generally, it provides a relatively small increase in Setup 8 compared to Setup 4. We believe that the reason behind this is that CE and ME(CHAR) competes with each other in representing the morphological information of the word. The reason that Setup 9 achieved higher performance compared to Setup 8 is probably because the missing roots in Setup 9 can be covered by CE combined with relatively lower complexity of ME(WOR).

We have also evaluated our model on text in Czech which is another morphologically rich language. To be able to compare our results, we used the CNEC 2.0 corpus in CoNLL format as other studies did (Konkol and Konopík, 2013). We chose to include only the ME(CHAR) setup for Czech because it gave good results both with and without character embeddings.

Lastly, we compare our best results with previous state-of-the-art in Table 2. The performance of (Seker and Eryiğit, 2012) without gazetteers is 89.55%, (Kuru et al., 2016) does not employ any external data and (Demir and Özgür, 2014) still relies on hand-crafted features despite exploiting word embeddings trained externally.

	F1-Measure	
Model	**Turkish**	**Czech**
(Kuru et al., 2016)	91.30	72.19
(Demir and Özgür, 2014)	91.85	75.61
(Seker and Eryiğit, 2012)	91.94	N/A
This work	**93.59**	**79.59**

Table 2: Comparison with previous work.

4 Conclusions

In this work, we demonstrated a new state-of-the-art system for Turkish and Czech named entity recognition using the model of (Lample et al., 2016). We introduced embedding configurations to understand the affect of different combinations of the morphological tags. Using these configurations, we showed that augmenting word representations with morphological embeddings improves the performance. However, the contribution of morphological embeddings seems to be subsumed by character embeddings in some of these configurations. Thus a thorough examination and comparison of character and morphological embeddings learned in this sense is required for further discussion.

References

Douglas E Appelt, Jerry R Hobbs, John Bear, David Israel, Megumi Kameyama, David Martin, Karen Myers, and Mabry Tyson. 1995. SRI International FASTUS system: MUC-6 test results and analysis. In *Proceedings of the 6th Conference on Message Understanding*. ACL, pages 237–248.

Bogdan Babych and Anthony Hartley. 2003. Improving machine translation quality with automatic named entity recognition. In *Proceedings of the 7th International EAMT Workshop on MT and Other Language Technology Tools, Improving MT through Other Language Technology Tools: Resources and Tools for Building MT*. ACL, pages 1–8.

Parminder Bhatia, Robert Guthrie, and Jacob Eisenstein. 2016. Morphological priors for probabilistic neural word embeddings. In *EMNLP*.

Andrew Eliot Borthwick. 1999. *A Maximum Entropy Approach to Named Entity Recognition*. Ph.D. thesis, New York University, New York, NY, USA.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing. In *Proceedings of the 25th International Conference on Machine Learning - ICML 08*. ACM. https://doi.org/10.1145/1390156.1390177.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research* 12(Aug):2493–2537.

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

Hakan Demir and Arzucan Özgür. 2014. Improving named entity recognition for morphologically rich languages using word embeddings. In *Machine Learning and Applications (ICMLA), 2014 13th International Conference on*. IEEE, pages 117–122.

Thomas G Dietterich. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation* 10(7):1895–1923.

Cícero Nogueira dos Santos and Bianca Zadrozny. 2014. Learning character-level representations for part-of-speech tagging. In *ICML*.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting of ACL*. ACL, pages 363–370.

Honglei Guo, Huijia Zhu, Zhili Guo, Xiaoxun Zhang, Xian Wu, and Zhong Su. 2009a. Domain adaptation with latent semantic association for named entity recognition. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the NAACL*. ACL, pages 281–289.

Jiafeng Guo, Gu Xu, Xueqi Cheng, and Hang Li. 2009b. Named entity recognition in query. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, pages 267–274.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9(8).

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991* .

Kevin Humphreys, Robert Gaizauskas, Saliha Azzam, Chris Huyck, Brian Mitchell, Hamish Cunningham, and Yorick Wilks. 1998. University of Sheffield: Description of the LaSIE-II system as used for MUC-7. In *Proceedings of the Seventh Message Understanding Conferences (MUC-7)*. ACL.

Jing Jiang and Cheng Xiang Zhai. 2007. Instance weighting for domain adaptation in NLP. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*. ACL, Prague, volume 7, pages 264–271.

Michal Konkol and Miloslav Konopík. 2013. Crf-based czech named entity recognizer and consolidation of czech ner research. In Ivan Habernal and Václav Matoušek, editors, *Text, Speech and Dialogue*, Springer Berlin Heidelberg, volume 8082 of *Lecture Notes in Computer Science*, pages 153–160.

Onur Kuru, Ozan Arkan Can, and Deniz Yuret. 2016. Charner: Character-level named entity recognition. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, Osaka, Japan, pages 911–921. http://aclweb.org/anthology/C16-1087.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the eighteenth international conference on machine learning, ICML*. volume 1, pages 282–289.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of NAACL-HLT 2016*. pages 260–270.

Matti Lankinen, Hannes Heikinheimo, Pyry Takala, Tapani Raiko, and Juha Karhunen. 2016. A character-word compositional neural language model for finnish. *CoRR* abs/1612.03266.

Changki Lee, Yi-Gyu Hwang, and Myung-Gil Jang. 2007. Fine-grained named entity recognition and relation extraction for question answering. In *Proceedings of the 30th Annual International*

ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR 07. ACM. https://doi.org/10.1145/1277741.1277915.

Ye Liu and Fuji Ren. 2011. Japanese named entity recognition for question answering system. In *2011 IEEE International Conference on Cloud Computing and Intelligence Systems*. IEEE. https://doi.org/10.1109/ccis.2011.6045098.

Thang Luong, Richard Socher, and Christopher D Manning. 2013. Better word representations with recursive neural networks for morphology. In *CoNLL*. pages 104–113.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. *arXiv preprint arXiv:1603.01354* .

Andrew McCallum and Wei Li. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL*. Association for Computational Linguistics, volume 4, pages 188–191.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Interspeech*. volume 2, page 3.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*. pages 3111–3119.

Kemal Oflazer. 1994. Two-level description of Turkish morphology. *Literary and Linguistic Computing* 9(2):137–148.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pages 1532–1543. http://www.aclweb.org/anthology/D14-1162.

Gökhan Akın Seker and Gülşen Eryiğit. 2012. Initial explorations on using CRFs for Turkish named entity recognition. In *Proceedings of COLING 2012*. The COLING 2012 Organizing Committee, Mumbai, India, pages 2459–2474. http://www.aclweb.org/anthology/C12-1150.

Qinlan Shen, Daniel Clothiaux, Emily Tagtow, Patrick Littell, and Chris Dyer. 2016. The role of context in neural morphological disambiguation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, Osaka, Japan, pages 181–191. http://aclweb.org/anthology/C16-1018.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ACL, Seattle, Washington, USA, pages 1631–1642. http://www.aclweb.org/anthology/D13-1170.

György Szarvas, Richárd Farkas, and András Kocsor. 2006. A multilingual named entity recognition system using boosting and C4.5 decision tree learning algorithms. In *International Conference on Discovery Science*. Springer, pages 267–278.

Gökhan Tür, Dilek Hakkani-Tür, and Kemal Oflazer. 2003. A statistical information extraction system for turkish. *Natural Language Engineering* 9(2):181–210.

Joseph Turian, Lev-Arie 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*. Association for Computational Linguistics, Uppsala, Sweden, pages 384–394. http://www.aclweb.org/anthology/P10-1040.

Dan Wu, Wee Sun Lee, Nan Ye, and Hai Leong Chieu. 2009. Domain adaptive bootstrapping for named entity recognition. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ACL, volume 3, pages 1523–1532.

Yang Xu and Jiawei Liu. 2017. Implicitly incorporating morphological information into word embedding. *CoRR* abs/1701.02481.

Zhilin Yang, Ruslan Salakhutdinov, and William W. Cohen. 2016. Multi-task cross-lingual sequence tagging from scratch. *CoRR* abs/1603.06270.

Reyyan Yeniterzi. 2011. Exploiting morphology in turkish named entity recognition system. In *Proceedings of the ACL 2011 Student Session*. Association for Computational Linguistics, Stroudsburg, PA, USA, HLT-SS '11, pages 105–110. http://dl.acm.org/citation.cfm?id=2000976.2000995.

Eray Yildiz, Caglar Tirkaz, H Bahadir Sahin, Mustafa Tolga Eren, and Omer Ozan Sonmez. 2016. A morphology-aware network for morphological disambiguation. In *30th AAAI Conference on Artificial Intelligence*.