Neural Networks for Featureless Named Entity Recognition in Czech

Jana Straková, Milan Straka, and Jan Hajič

Charles University in Prague, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics,
Malostranské náměstí 25,
118 00 Prague, Czech Republic
{strakova,straka,hajic}@ufal.mff.cuni.cz
http://ufal.mff.cuni.cz

Abstract. We present a completely featureless, language agnostic named entity recognition system. Following recent advances in artificial neural network research, the recognizer employs parametric rectified linear units (PReLU), word embeddings and character-level embeddings based on gated linear units (GRU). Without any feature engineering, only with surface forms, lemmas and tags as input, the network achieves excellent results in Czech NER and surpasses the current state of the art of previously published Czech NER systems, which use manually designed rule-based orthographic classification features. Furthermore, the neural network achieves robust results even when only surface forms are available as input. In addition, the proposed neural network can use the manually designed rule-based orthographic classification features and in such combination, it exceeds the current state of the art by a wide margin.

Key words: neural networks, named entity recognition, Czech, word embeddings, character-level embeddings, parametric rectified linear unit (PReLU), gated linear unit (GRU)

1 Introduction

Recent years have seen a dramatic progress in the field of artificial neural networks. The publication of word embeddings [1] opened reliable and computationally affordable ways of using tokens as classification features in artificial neural networks. For morphologically rich languages, word embeddings appear rather too coarse for many tasks, especially those where the inner structure of the word such as prefixes and suffixes, is crucial. Therefore, the ideas go further in publication of character-level embeddings [2], which recently improved the state of the art in POS-tagging [3]. One of the advantages of word- and character-level embeddings is that they are learned automatically from large raw corpora.

Another paradigm-changing publication introduces the long short-term memory units (LSTMs, [4]). In simple words, LSTMs are specially shaped units of artificial neural networks designed to process whole sequences. LSTMs have been

shown to capture non-linear and non-local dynamics in sequences [4] and have been used to obtain many state-of-the-art results in sequence classification [3,5].

Recently, a gated linear unit (GRU) was proposed by [6] as an alternative to LSTM, and was shown to have similar performance, while being less computationally demanding.

In this work, we use artificial neural networks employing parametric rectified linear units (PReLU), word embeddings and character-level embeddings based on gated linear units (GRU). We describe our methodology in Section 3. We report our results and discussion in Section 4 and we conclude in Section 5.

2 Related Work

Czech named entity recognition (NER) has become a well-established field. Following the publication of the Czech Named Entity Corpus [7,8], a selection of named entity recognizers for Czech has been published: [8,9,10,11,12]and even a publicly available Czech named entity recognition exists (NameTag,[1] [13]).

All these works use manually selected rule-based orthographic classification features, such as first character capitalization, existence of special characters in the word, regular expressions designed to reveal particular named entity types. Also gazetteers are extensively utilized. The authors employed a wide selection of machine learning techniques (decision trees [7], SVMs [8], maximum entropy classifier [9], CRFs [10]), clustering techniques [12] and stemming approaches [11].

The contribution of our work is that we use artificial neural networks with parametric rectified linear units, word embeddings and character-level embeddings, which do not need manually designed classification features or gazetteers, and still surpass the current state of the art.

In [14], the authors present a semi-supervised learning approach based on neural networks for Czech and Turkish NER utilizing word embeddings [1], but there are some differences in the neural network design and in classification features used. Instead regularized averaged perceptron, we use parametric rectified linear units, character-level embeddings and dropout. The NER system in [14] does not use morphological analysis, it is therefore similar to our experiments with only surface forms as input. However, the system does use "type information of the window c_i, i.e. is-capitalized, all-capitalized, all-digits, ..." etc. Our system surpasses these results even without using such features.

English named entity recognition has a successful tradition in computational linguistics and the state of the art [15] has recently been pushed forward by [16,17,18,5]. We present a comparison with these works in Section 4. The most similar to our proposed design is [5], which was accepted to NAACL 2016 the exact month of this paper submission. The authors propose a very similar network with LSTMs, word embeddings and character-level embeddings. However, while we classify each word separately and use Viterbi to perform the final decoding, [5] employs LSTMs combined with CRF layer to decode whole sentences, which brings a determining advantage over our framework as we show in Section 4.

[1] `http://ufal.mff.cuni.cz/nametag`

3 Methodology

We conduct our experiments on all available Czech NER corpora, so that we are able to compare with all available related work in Czech NER: Czech Named Entity Corpus (CNEC) 1.0 [7,8], CNEC 2.0,[2], CoNLL-based Extended CNEC 1.1 [10], CoNLL-based Extended CNEC 2.0.[3].

The named entities in CNEC 1.0 and 2.0 are hierarchically organized in two hierarchies (fine-grained "types" and coarse "supertypes", [8]), may be nested and labeled with more than one label.[4]

CoNLL-based Extended CNEC 1.1 and 2.0 are based on the respective original CNEC corpora, but they use only the coarser 7 classes, are flattened and assume that entities are non-nested and labeled with one label.

For comparison with the English state of the art, we evaluated our NER system on CoNLL-2003 shared task dataset [19]. In this task, four classes are predicted: PER (person), LOC (location), ORG (organization) and MISC (miscellaneous). The named entities are non-nested, non-overlapping and annotated with exactly one label.

In case of the original CNEC 1.0 and CNEC 2.0, we present results for both fine-grained and coarse-grained classes hierarchy ("types" and "supertypes", [8]) and we evaluate our results with the script provided with the corpora, which computes F-measure of selected types [8].

In case of the CoNLL-based Extended CNEC 1.1 and 2.0, we present results for the 7 classes present in these corpora and evaluate our results with the standard CoNLL evaluation script `conlleval`.

Similarly, the English CoNLL-2003 dataset is evaluated with CoNLL evaluation script `conlleval`.

3.1 The Network Classifier

For each word (and its context), we compute the probability distribution of labeling this word with BILOU-encoded [15] named entities. We then determine the best consistent assignment of BILOU-encoded entities to the words in the sentence using the Viterbi algorithm.

We compute the probability distribution for each word using an artificial neural network. The input layer consists of representations of surface forms (and optionally lemmas, tags, characters, character-level embeddings and classification features) of the word and W previous and W following words. The input layer is connected to a hidden layer of parametric rectified linear units [20] and the hidden layer is connected to the output layer which is a softmax layer producing probability distribution for all possible named entity classes in BILOU encoding.

We represent each word using a combination of the following:

[2] `http://ufal.mff.cuni.cz/cnec/cnec2.0`

[3] `http://home.zcu.cz/~konkol/cnec2.0.php`

[4] Our system learns and predicts only outermost entities and is thus penalized for every misted nested named entity during evaluation.

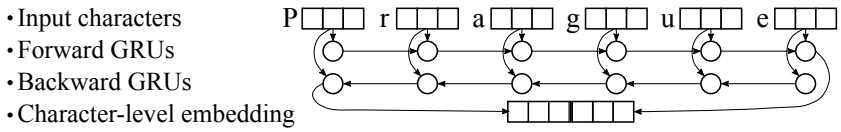

Fig. 1. Neural network for character-level embedding computation.

- *word embedding*: Word embeddings are vector representations of low dimension [1]. We generated the word embeddings using `word2vec` [1] and we chose the Skip-gram model with negative sampling.[5]
- *character-level embedding*: To overcome drawbacks of word embeddings (embeddings for different words are independent; unknown words cannot be handled), several orthography aware models have been proposed [2,3], which compute word representation from the characters of the word.
  We hypothesized that character-level embeddings such as published in [3] have the potential to increase the performance of Czech NER system. Our assumption was that Czech as a morphologically rich language would benefit from character-level embeddings rather than word embeddings especially in cases where no morphological analysis is available.
  We use bidirectional GRUs [6,21] in line with [3]: we represent every Unicode character with a vector of C real numbers, and we use GRUs to compute two outputs, a sequence of word characters and a sequence of reversed word characters, and we then concatenate the two outputs, as shown in Fig. 3.1.
- *prefix and suffix*: For comparison with character-level embeddings, we also include "poor man's" character-level embeddings – we encode first two and last two characters encoded as one-hot vectors. We hypothesize that character-level embeddings as a more sophisticated means should perform better.
- *tag*: We encode part-of-speech tags as one-hot vectors.
- *manually designed classification features*: We also publish a combination of our neural network framework with traditional manually designed rule-based orthographic classification features. We use quite a limited set of classification features inspired by [9]: capitalization information, punctuation information, number information and Brown clusters [22]. We do not use gazetteers, context aggregation, prediction history nor two-stage decoding.

The network is trained with AdaGrad [23] and we use dropout [24] on the hidden layer. We implemented our neural network in Torch7 [25], a scientific computing framework with wide support for machine learning algorithms.

We tuned most of the hyperparameters on development portion of CNEC 1.0 and used them for all other corpora. Notably, we utilize window size $W = 2$, hidden layer of 200 nodes, dropout 0.5, minibatches of size 100 and learning rate 0.02 with decay. We tune the dimension C of the character-level embeddings for every corpus separately, choosing either 32 or 64. All reported experiments use

[5] We used the following options: `-cbow 0 -window 5 -negative 5 -iter 1`

an ensemble of 5 networks, each using different random seed, with the resulting distributions being an average of individual networks distributions. The training of a single network took half a day on a single CPU to stabilize performance on development data. During evaluation of testing data, we add the development data to the training data, a technique proposed in context of NER by [15].

We trained the word embeddings of dimension 200 on English Gigaword Fifth Edition corpus and on Czech SYN [26]. We also lemmatized the corpora with MorphoDiTa [13] in order to train the lemma embeddings.

4 Results and Discussion

We present two groups of experiments with low and high complexity depending on the available network input: experiments where only surface forms were used, a putatively more difficult task as no linguistic knowledge is available to the NER system; and experiments with morphologically analyzed and POS-tagged text. We automatically generate lemmas and POS-tags from surface forms with MorphoDiTa [13], an open source tagger and lemmatizer.

Table 1 presents all results of this work. Our baseline is an artificial neural network with only surface forms encoded as word embeddings. We then add more computational complexity to the network: WE stands for word embeddings of forms and lemmas, CLE stands for character-level embeddings of forms and lemmas, 2CH stands for first two and last two characters of forms, lemmas and POS tags, and CF stands for experiments with traditional classification features.

4.1 Experiments with Surface Forms in Czech

This group of experiments dealt with situations when only surface forms are available as input. Since most of the previous literature heavily depends on manually selected language-dependent features, as well as gazetteers and more or less linguistically motivated variants of lemmatization of stemming, the only work to be directly compared with is [14]. The authors of [14] use a similar, semi-supervised neural network based approach. Their final system, which uses word embeddings, capitalization and punctuation information, prefixes, suffixes, context aggregation and prediction history, achieves CoNLL F-measure 75.61 for CoNLL-based Extended CNEC 1.1. We surpass these results with CoNLL F-measure 76.72, using only word embeddings, character-level embeddings and first two and last two characters. If the traditional features are added, we even achieve CoNLL F-measure 78.21.

4.2 Experiments with Lemmas and POS Tags in Czech

Table 1 presents a comparison with related work on all available Czech NER corpora. The row denoted *f,l,t+WE+CLE+2CH+CF* is our best setting, including manually selected classification features. Our proposed network clearly exceeds

Experiment/Related Work	Corpus						
	Original CNEC 1.0		Original CNEC 2.0		Extended CNEC 1.1	Extended CNEC 2.0	English CoNLL-2003
	Types	Supt.	Types	Supt.	Classes	Classes	Classes
f+WE (`baseline`)	63.24	69.61	63.33	68.87	63.48	63.91	67.99
f+CLE	71.43	76.13	70.50	75.80	69.59	70.06	82.65
f+WE+2CH	69.73	74.49	69.44	74.31	75.15	74.36	79.40
f+WE+CLE	73.30	78.11	73.10	77.89	73.33	73.80	84.08
f+WE+CLE+2CH	73.71	78.32	72.81	77.87	76.72	77.18	84.29
f+WE+CLE+2CH+CF	73.73	78.50	72.91	77.65	78.21	78.20	86.06
f,l,t+WE	80.07	83.21	77.45	80.92	78.42	78.18	87.92
f,l,t+CLE	75.63	80.88	74.38	79.85	75.32	76.02	83.70
f,l,t+WE+2CH	80.46	83.85	78.32	82.09	79.68	79.48	89.37
f,l,t+WE+CLE	80.64	84.06	78.62	82.48	80.11	80.41	89.74
f,l,t+WE+CLE+2CH	80.92	84.18	78.63	82.41	**80.88**	**80.79**	89.71
f,l,t+WE+CLE+2CH+CF	**81.20**	**84.68**	**79.23**	**82.78**	80.73	80.73	**89.92**
Kravalová et al., 2009 [8]	*68.00*	*71.00*	–	–	–	–	–
Konkol et al., 2013 [10]	–	*79.00*	–	–	*74.08*	–	*83.24*
Straková et al., 2013 [9]	*79.23*	*82.82*	–	–	–	–	–
Konkol et al., 2014 [11]	–	–	–	–	*74.23*	*74.37*	–
Demir at al., 2014 [14]	–	–	–	–	*75.61*	–	–
Konkol et al., 2015 [12]	–	–	–	–	*74.08*	–	*89.44*
Ratinov et al., 2009 [15]	–	–	–	–	–	–	*90.80*
Lin et al., 2009 [16]	–	–	–	–	–	–	*90.90*
Chiu et al., 2015 [17]	–	–	–	–	–	–	*90.77*
Luo et al., 2015 [18]	–	–	–	–	–	–	*91.20*
Lample et al., 2016 [5]	–	–	–	–	–	–	*90.94*

Table 1. Experiment results and comparison with related work. Columns denote corpora, rows our experiments or related work. First group of rows describes our experiments with surface forms only (*f*), second group our experiments with forms, lemmas and POS-tags (*f,l,t*). *WE* stands for word embeddings, *CLE* for character-level embeddings, *2CH* for first two and last two characters, *CF* for traditional classification features. Third group of rows describes related work in Czech NER, and fourth group related work in English NER.

the current state of the art on all Czech corpora in measures selected by the authors of the respective literature.

We shall now focus our discussion on featureless neural networks. Our system exceeds the current Czech state of the art solely with automatically obtained word embeddings (see row *f,l,t+WE* in Table 1), without requiring manually designed rule-based orthographic features, gazetteers, context aggregation, prediction history or two-stage decoding. The effect is even stronger with character-level embeddings and optionally first two and last two characters.

4.3 English Experiments

Our best result (row *f,l,t+WE+CLE+2CH+CF*) is F-measure 89.92, which is near the English state of the art. A work most similar to ours, [5], also proposed neural network architecture with word embeddings and character-level embeddings. Nevertheless, in [5] sentence-level decoding using bidirectional LSTMs with additional CRF layer is used, while our framework decodes the entities using Viterbi algorithm on probability distributions of named entity classes.

5 Conclusions

We presented an artificial neural network based NER system which achieves excellent results in Czech NER and near state-of-the-art results in English NER without manually designed rule-based orthographic classification features, gazetteers, context aggregation, prediction history or two-stage decoding. Our proposed architecture exceeds all known Czech published results only with forms, lemmas and POS tags encoded as word embeddings and achieves even better results in combination with character-level embeddings, prefixes and suffixes. Finally, it surpasses the current state of the art of Czech NER in combination with traditional classification features by a wide margin. The proposed neural network also yields very robust results without morphologic analysis or POS-tagging, when only surface forms are available. As our future work, we plan to improve our decoding in line with [5].

Acknowledgments

This work has been partially supported and has been using language resources and tools developed, stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071). This research was also partially supported by SVV project number 260 333.

References

1. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems 26. Curran Associates, Inc. (2013) 3111–3119
2. Santos, C.D., Zadrozny, B.: Learning character-level representations for part-of-speech tagging. In: Proceedings of the 31st International Conference on Machine Learning, JMLR Workshop and Conference Proceedings (2014) 1818–1826
3. Ling, W., Luís, T., Marujo, L., Astudillo, R.F., Amir, S., Dyer, C., Black, A.W., Trancoso, I.: Finding Function in Form: Compositional Character Models for Open Vocabulary Word Representation. CoRR **abs/1508.02096** (2015)
4. Hochreiter, S., Schmidhuber, J.: Long Short-Term Memory. Neural Comput. **9**(8) (November 1997) 1735–1780
5. Lample, G., Ballesteros, M., Kawakami, K., Subramanian, S., Dyer, C.: Neural Architectures for Named Entity Recognition. CoRR **abs/1603.01360v1** (2016) To appear at NAACL 2016.
6. Cho, K., van Merrienboer, B., Bahdanau, D., Bengio, Y.: On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. CoRR **abs/1409.1259** (2014)
7. Ševčíková, M., Žabokrtský, Z., Krůza, O.: Named entities in Czech: annotating data and developing NE tagger. In: Proceedings of the 10th international conference on Text, speech and dialogue. TSD'07, Springer-Verlag (2007) 188–195
8. Kravalová, J., Žabokrtský, Z.: Czech named entity corpus and SVM-based recognizer. In: Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration. NEWS '09, ACL (2009) 194–201

9. Straková, J., Straka, M., , Hajič, J.: A New State-of-The-Art Czech Named Entity Recognizer. In: Text, Speech, and Dialogue: 16th International Conference, Berlin, Heidelberg, Springer Berlin Heidelberg (2013) 68–75
10. Konkol, M., Konopík, M.: CRF-based Czech named entity recognizer and consolidation of Czech NER research. In: Text, Speech, and Dialogue, Springer Berlin Heidelberg (2013) 153–160
11. Konkol, M., Konopík, M.: Named entity recognition for highly inflectional languages: effects of various lemmatization and stemming approaches. In: Text, Speech and Dialogue, Springer International Publishing (2014) 267–274
12. Konkol, M., Brychcín, T., Konopík, M.: Latent semantics in named entity recognition. Expert Systems with Applications **42**(7) (2015) 3470–3479
13. Straková, J., Straka, M., Hajič, J.: Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Baltimore, Maryland, ACL (June 2014) 13–18
14. Demir, H., Özgür, A.: Improving Named Entity Recognition for Morphologically Rich Languages Using Word Embeddings. In: Machine Learning and Applications (ICMLA), 2014 13th International Conference on. (Dec 2014) 117–122
15. Ratinov, L., Roth, D.: Design challenges and misconceptions in named entity recognition. In: CoNLL '09: Proceedings of the Thirteenth Conference on Computational Natural Language Learning, ACL (2009) 147–155
16. Lin, D., Wu, X.: Phrase clustering for discriminative learning. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2, Association for Computational Linguistics (2009) 1030–1038
17. Chiu, J.P.C., Nichols, E.: Named Entity Recognition with Bidirectional LSTM-CNNs. CoRR **abs/1511.08308** (2015)
18. Luo, G., Huang, X., Lin, C.Y., Nie, Z.: Joint Named Entity Recognition and Disambiguation. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, ACL (2015) 879–888
19. Tjong Kim Sang, E.F., De Meulder, F.: Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In: Proceedings of CoNLL-2003, Edmonton, Canada (2003) 142–147
20. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. CoRR **abs/1502.01852** (2015)
21. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional lstm and other neural network architectures. Neural Networks (2005) 5–6
22. Brown, P.F., deSouza, P.V., Mercer, R.L., Pietra, V.J.D., Lai, J.C.: Class-based n-gram models of natural language. CL **18**(4) (December 1992) 467–479
23. Duchi, J., Hazan, E., Singer, Y.: Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. J. Mach. Learn. Res. **12** (July 2011) 2121–2159
24. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research **15** (2014) 1929–1958
25. Collobert, R., Kavukcuoglu, K., Farabet, C.: Torch7: A Matlab-like environment for machine learning. In: BigLearn, NIPS Workshop. (2011)
26. Hnátková, M., Křen, M., Procházka, P., Skoumalová, H.: The SYN-series corpora of Written Czech. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, ELRA (May 2014)