Porjazovski, Dejan; Leinonen, Juho; Kurimo, Mikko

Named Entity Recognition for Spoken Finnish

Published in:
AI4TV 2020 - Proceedings of the 2nd International Workshop on AI for Smart TV Content Production, Access and Delivery

DOI:
10.1145/3422839.3423066

Published: 12/10/2020

Document Version
Peer reviewed version

Please cite the original version:
Porjazovski, D., Leinonen, J., & Kurimo, M. (2020). Named Entity Recognition for Spoken Finnish. In *AI4TV 2020 - Proceedings of the 2nd International Workshop on AI for Smart TV Content Production, Access and Delivery* (pp. 25-29). ACM. https://doi.org/10.1145/3422839.3423066

Named Entity Recognition for Spoken Finnish

In this paper we present a Bidirectional LSTM neural network with a Conditional Random Field layer on top, which utilizes word, character and morph embeddings in order to perform named entity recognition on various Finnish datasets. To overcome the lack of annotated training corpora that arises when dealing with low-resource languages like Finnish, we tried a knowledge transfer technique to transfer tags from Estonian dataset. On the human annotated in-domain Digitoday dataset, out system achieved F1 score of 84.73. On the out-of-domain Wikipedia set we got F1 score of 67.66. In order to see how well the system performs on speech data, we used two datasets containing automatic speech recognition outputs. Since we do not have true labels for those datasets, we used a rule-based system to annotate them and used those annotations as reference labels. On the first dataset which contains Finnish parliament sessions we obtained F1 score of 42.09 and on the second one which contains talks from Yle Pressiklubi we obtained F1 score of 74.54.

CCS Concepts: • **Computing methodologies** → **Artificial intelligence**; **Natural language processing**;

Additional Key Words and Phrases: named entity recognition, speech recognition, low-resource

ACM Reference Format:
. 2018. Named Entity Recognition for Spoken Finnish. 1, 1 (August 2018), 8 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Named entity recognition (NER) is a natural language processing (NLP) task, in which the system aims to find entities in a text and classify them to predefined categories. The categories can vary based on the domain in which they are going to be used but some of the most common categories include: person, organization, product, location and date. NER is an integral part in larger areas such as information retrieval, question answering, machine translation and text summarization. It is a difficult problem because in most languages there is little annotated data available, especially in specific domains such as: chemistry, biology and medical fields. Entity ambiguity is another challenge that the system needs to deal with. For example "Facebook" can refer to both company and product, depending on the context it appears in.

In the past, researchers relied on hand-crafted features and gazetteers for solving this task, which requires in-domain knowledge [3, 9]. With the rise of machine learning (ML), researchers have tried different ML techniques in order to solve this task. The most popular approach is using Conditional Random Fields (CRFs) [11] which has been successfully applied to various NER tasks [14, 18].

With the increase of the computational power, deep neural networks have become more appealing, especially because they help to alleviate the need of domain experts and hand-crafted features. Recurrent neural networks, especially the Long Short-Term Memory (LSTM) [7] have been well suited for sequential tasks due to their ability to store information about sequences. Deep neural network architectures have been successfully applied in various NER tasks where they

Author's address:

Manuscript submitted to ACM

outperform CRF based models [1]. These systems used word embeddings as input features to the network. In 2015, a new neural network approach was proposed, which uses a CRF layer on top of a Bidirectional LSTM (BiLSTM). This model outperformed the previous neural network architectures and achieved the state-of-the-art results on the CoNLL 2013 dataset [8].

Although these approaches work well for most languages, frequent occurrences of inflections, derivations and compounding in morphologically rich languages makes their vocabulary large and increases the number of out-of-vocabulary (OOV) words. In order to overcome that issue, subword units such as characters or morphs have been proposed to replace words [6]. Character-level LSTM models have been applied on various languages and have been shown to give competitive results [10]. Furthermore, combining word and character representations has given an improvement over the existing models [13, 17]. Segmenting the words into morphs has been shown to improve the performance of the language models and reduce the out-of-vocabulary words by constructing the unseen words from morphs [2, 19].

Large vocabulary is not the only issue that arises when dealing with Finnish language. Lack of annotated data puts the Finnish language in a low-resource category. This constraint causes difficulties for training a NER system, especially when the annotated data is in a specific domain. In an attempt to overcome this limitation, different knowledge-transfer techniques have been proposed, which try to transfer tags from the source to the target language, in order to enrich the annotated corpora. [4, 23].

Another challenge that arises when doing NER is when we are dealing with unstructured data, such as an output of an automatic speech recognition (ASR) system. Named entities are often capitalized, so the system relies on the capitalization in order to detect the entities, which causes problems for ASR output, where capitalization is neglected.

In this paper we propose a bidirectional LSTM-CRF architecture that utilizes words, characters and morphs in order to achieve competitive results in NER for Finnish language. Moreover, we are going to explore different ways of improving the performance of the system on ASR output. In order to deal with the low-resource limitations, we experimented with knowledge transfer from Estonian language using multilingual word embeddings for Finnish and Estonian languages, aligned in a single vector space.

2 DATA

We used the Digitoday dataset to train the model. The dataset was collected and provided by [16]. It consists of online Finnish technological news articles. There are 953 articles and 193,742 word tokens in the dataset. Since the articles are from one domain, the authors also provided a Wikipedia test for evaluating the system on out-of-domain data. Both datasets are annotated using the BIO annotation scheme [15]. The Wikipedia test set consists of 83 articles and 49,752 word tokens. The dataset consists of 6 named entity classes:

- PERSON (PER)
- LOCATION (LOC)
- ORGANIZATION (ORG)
- PRODUCT (PROD)
- EVENT (EVENT)
- DATE (DATE)

The class distribution of Digitoday and Wikipedia datasets is presented in Table 1. Both datasets provide top-level and nested-level entities. In our experiments we used only the top-level entities.

Table 1. Class distribution in Digitoday and Wikipedia datasets.

Class	Count Digitoday	Count Wikipedia
ORG	15445	1821
LOC	4159	1427
PER	6517	2492
DATE	3685	1862
PRO	11655	2135
EVENT	569	362
TOTAL	42030	10099

Table 2. Class distribution in Parliament and Yle Pressiklubi datasets.

Class	Parliament	Yle Pressiklubi
PER	104	1350
LOC	54	601
ORG	/	327
TOTAL	158	2278

Table 3. Class distribution in Estonian dataset.

Class	Count
PER	12154
LOC	3508
ORG	9424
TOTAL	25086

In order to test how well the system performs in an ASR setting, we used two datasets of ASR outputs. The first one contains Finnish parliament sessions and the second one contains talks from Yle Pressiklubi television show. Using a commercial rule-based system, we managed to obtain two NER tags for the parliament sessions and three tags for the Pressiklubi dataset. The Parliament dataset is in lowercase and without punctuation, whereas the Yle Pressiklubi dataset is re-capitalized. The class distribution for the ASR datasets is presented in Table 2. In Table 3 we can see the class distribution of the Estonian dataset that we used to transfer tags to Finnish and make the dataset less domain biased.

3 METHODS

In this section we present our architecture for NER, which utilizes word, character and morph representations. For agglutinative languages like Finnish, which have a rich vocabulary, the number of OOV words increases, which has an impact on the performance of the model. In order to mitigate this, besides the standard word embeddings, we augmented our model with morph and character representations of the words. To obtain the morphs, we used the Morfessor toolkit [20]. The architecture is depicted in Figure 1.

As we can see from the figure, word, character and morph embeddings are processed through separate BiLSTMs. The outputs of the BiLSTMs are concatenated in order to get a single representation. The concatenated outputs then go through a highway layer, which is followed by a fully connected layer. At the end, the output of the fully connected layer goes through a CRF layer, which produces tag probabilities.

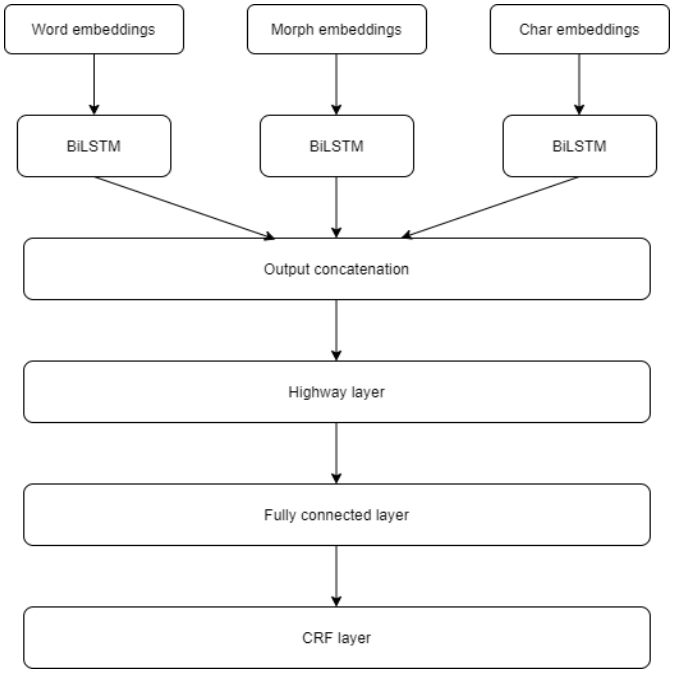

Fig. 1. BiLSTM-CRF model that utilizes char, morph and word embeddings.

4 EXPERIMENTS

As described in section 2, we use the Digitoday dataset which contains technological articles, as well as the Wikipedia test set and the ASR outputs for testing the system on out-of-domain data. For the Digitoday and Wikipedia evaluations, we trained our system using the Digitoday train set and for the ASR evaluations we used the whole Digitoday dataset along with the Wikipedia test set for training.

In order to make a distinction between the first and the last word in a sentence and the rest of the words, we added "<*start*>" and "<*end*>" tokens to each sentence. For the morph-based subword modeling, we added boundary markers to enforce restrictions on the generated output. Different ways of adding markers enforce different restrictions. Some common types of markers are: "<*w*>", "<*m+*>", "<*+m*>", "<*+m+*>". In our experiments we used the "<*+m+*>" style marker since it has been shown to give best results for Finnish language modeling in ASR [21]. For example, the word 'mobiilikäyttöjärjestelmä' would be segmented as 'mobiili+ +käyttö+ +järjestelmä'.

The architecture has 2 BiLSTM layers and 4 highway layers. The embedding dimensions of words, chars and morphs are 300, 100, 100 respectively for the BiLSTM networks. The hidden sizes are 300, 75, 75 for words, chars and morphs. A dropout of 0.5 is added to the final BiLSTM outputs and 0.2 for each layer except for the last. After the highway layer, we added a dropout probability of 0.7. For training the model we used a batch size of 128 and RAdam optimizer [12] with learning rate of 0.001. All of the hyperparameters were chosen based on internal experiments that we did on the development set.

As baseline models we used a rule based system called FiNER and a neural network architecture called GÜNGÖR-NN [16]. The GÜNGÖR-NN architecture is described in more detail in [5]. In order to see the system's performance on ASR

Table 4. Overall micro average precision, recall and F1 scores for the top-level entities of Digitoday test set.

architecture	precision	recall	F1
FiNER	90.41	83.51	86.82
GÜNGÖR-NN	83.59	85.62	84.59
word+char+morph-LSTM	85.52	83.74	84.62
word+char+morph-LSTM+transfer	85.27	84.19	84.73

output, we used Finnish parliament sessions as well as the Yle Pressiklubi television show, which were decoded using a commercial ASR system. The datasets were annotated by named entity tags given by a rule-based system. We used those tags as the reference labels.

Doing NER on an ASR output has many challenges, such as recognition errors and missing capitalization and punctuation. When evaluating the model on the Parliament dataset, we decided to remove capitalization and punctuation from the training data, so that the system would learn in the ASR setting better.

Another issue that we faced was the out-of-domain problem, just like when testing our system on the Wikipedia dataset. To alleviate that problem we used knowledge transfer technique as described in the previous section. Because some of the tag translations were not very accurate, we used thresholding to keep only the translations that have high nearest neighbor candidate score in the target language. We did multiple experiments on the Digitoday dev set and found that a threshold value of 0.6 yields best results. Since person names and location names are almost the same in Finnish and Estonian, we kept them as they are in the Estonian and just transferred them to Finnish. This approach gave us an improvement over translating them as we did with the other entities.

5 RESULTS

In this section we present the results obtained from the proposed BiLSTM-CRF architecture and compare them with the rule-based and neural baseline models. We also provide the results obtained for the ASR outputs. Additionally, we will show how much improvement did the knowledge transfer method give. We used the micro F1 score evaluation metric [22] in all the experiments.

The final results for the Digitoday dataset are presented in Table 4 and for the out-of-domain Wikipedia test set in Table 5. In Table 6 we can see how well our model performs on the Parliament and Yle Pressiklubi datasets, annotated by the rule-based system.

Since the Parliament dataset is lowercased and without punctuation, during training, we simulated the same scenario and trained the model in that setting, which resulted in significant improvement. The results for the Parliament dataset when the training data is kept as it is (with capitalization and punctuation) is shown in Table 7.

To see how well our model agrees with the rule-based system, we evaluated the system only on entities that were found by that system. The results for the Parliament and Yle Pressiklubi datasets are shown in Table 8.

At the end, we manually annotated 50 sentences from both ASR datasets in order to see how well the system performs on gold standard data. The results from the manually annotated Parliament and Yle Pressiklubi datasets are presented in Table 9.

Table 5. Overall micro average precision, recall and F1 scores for the top-level entities of Wikipedia test set.

architecture	precision	recall	F1
FiNER	85.17	72.47	78.31
GÜNGÖR-NN	62.98	55.89	59.22
word+char+morph-LSTM	71.34	56.38	62.98
word+char+morph-LSTM+transfer	74.55	61.93	67.66

Table 6. Overall micro average precision, recall and F1 scores for the Parliament and Yle Pressiklubi datasets.

	Parliament data			Yle Pressiklubi data		
TAG	precision	recall	F1	precision	recall	F1
PER	46.11	89.25	60.81	80.00	85.71	82.76
LOC	14.53	69.39	24.03	76.92	86.96	81.63
ORG	/	/	/	55.56	26.79	36.14
avg	28.26	82.39	42.09	76.25	72.91	74.54

Table 7. Overall micro average precision, recall and F1 scores for the Parliament dataset, trained without removing capitalization and punctuation.

TAG	precision	recall	F1
PER	72.73	8.60	15.38
LOC	19.57	18.37	18.95
avg	29.82	11.97	17.09

Table 8. Overall micro average precision, recall and F1 scores for the Parliament and Yle Pressiklubi datasets, comparing only entities found by the rule-based system.

	Parliament data			Yle Pressiklubi data		
TAG	precision	recall	F1	precision	recall	F1
PER	98.81	89.25	93.79	85.04	85.71	85.38
LOC	100.00	69.39	81.93	89.55	86.96	88.24
ORG	/	/	/	78.95	26.79	40.00
avg	99.15	82.39	90.00	85.92	72.91	78.88

6 ANALYSIS OF THE RESULTS

From Table 4 we can see that when we added transferred tags from Estonian language, we gained a slight boost in the F1 score. Our model achieved F1 score of 84.73, which is slightly better than the GÜNGÖR-NN architecture. Still, our system performed worse than the rule-based FiNER system, which achieved F1 score of 86.82.

In Table 5 we can see the results for the Wikipedia test set. On this out-of-domain dataset, the knowledge transfer technique improved the F1 score from 62.98 to 67.66. Compared to the GÜNGÖR-NN architecture our system did far better but it still falls behind compared to the FiNER system. From the results in Tables 4 and 5 we can see that transferring tags from Estonian had bigger impact on the out-of-domain Wikipedia set than on the Digitoday test set. We can also observe that neural network architectures suffer more from out-of-domain data but our architecture still performs better than the GÜNGÖR-NN.

If we compare the results presented in Table 6, we can see that our systems has low precision for the Parliament data when evaluated against the rule-based system annotations. The reason is that our system is able to find more entities

Table 9. Overall micro average precision, recall and F1 scores for the manually annotated Parliament and Yle Pressiklubi datasets.

	Parliament data			Yle Pressiklubi data		
TAG	precision	recall	F1	precision	recall	F1
PER	91.43	84.21	87.67	91.11	85.42	88.17
LOC	77.27	80.95	79.07	84.62	84.62	84.62
ORG	/	/	/	100.00	32.14	48.65
avg	85.96	83.05	84.48	90.00	70.59	79.12

than the rule-based system and since those entities are not present in the annotations obtained by that system, we get high number of false positives.

When comparing only with the entities found by the rule-based system, we can see that our system agrees with the rule-based system almost all the time, which results in high precision.

When evaluated on the manually annotated data, we can see that our system achieves relatively good results.

7 CONCLUSION

In this paper we showed that our system which incorporates word, character and morph representations achieves competitive results on Digitoday dataset. Furthermore, we saw that transferring tags from Estonian language using multilingual embeddings significantly improved the results on the out-of-domain Wikipedia test set.
Additionally, we evaluated our system on two ASR output datasets, where one of them did not have capitalization and punctuation, which caused difficulties for our system. In order to mitigate those difficulties, we converted our training set to lowercase and removed the punctuation in order to simulate ASR setting, which yielded significant improvement.

REFERENCES

[1] 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 (2011), 2493–2537.

[2] Mathias Creutz, Teemu Hirsimäki, Mikko Kurimo, Antti Puurula, Janne Pylkkönen, Vesa Siivola, Matti Varjokallio, Ebru Arisoy, Murat Saraçlar, and Andreas Stolcke. 2007. Morph-based speech recognition and modeling of out-of-vocabulary words across languages. *ACM Transactions on Speech and Language Processing (TSLP)* 5, 1 (2007), 3.

[3] Dimitra Farmakiotou, Vangelis Karkaletsis, John Koutsias, George Sigletos, Constantine D Spyropoulos, and Panagiotis Stamatopoulos. 2000. Rule-based named entity recognition for Greek financial texts. In *Proceedings of the Workshop on Computational lexicography and Multimedia Dictionaries (COMLEX 2000)*. 75–78.

[4] Xiaocheng Feng, Xiachong Feng, Bing Qin, Zhangyin Feng, and Ting Liu. 2018. Improving Low Resource Named Entity Recognition using Cross-lingual Knowledge Transfer.. In *IJCAI*. 4071–4077.

[5] Onur Güngör, Suzan Üsküdarlı, and Tunga Güngör. 2018. Improving Named Entity Recognition by Jointly Learning to Disambiguate Morphological Tags. *arXiv preprint arXiv:1807.06683* (2018).

[6] Teemu Hirsimaki, Janne Pylkkonen, and Mikko Kurimo. 2009. Importance of high-order n-gram models in morph-based speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing* 17, 4 (2009), 724–732.

[7] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

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

[9] Jun'ichi Kazama and Kentaro Torisawa. 2008. Inducing gazetteers for named entity recognition by large-scale clustering of dependency relations. In *proceedings of ACL-08: HLT*. 407–415.

[10] 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*. 911–921.

[11] John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. (2001).

[12] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. 2019. On the Variance of the Adaptive Learning Rate and Beyond. *arXiv preprint arXiv:1908.03265* (2019).

[13] Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354* (2016).

[14] 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 seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*. Association for Computational Linguistics, 188–191.

[15] Lance A Ramshaw and Mitchell P Marcus. 1999. Text chunking using transformation-based learning. In *Natural language processing using very large corpora*. Springer, 157–176.

[16] Teemu Ruokolainen, Pekka Kauppinen, Miikka Silfverberg, and Krister Lindén. 2019. A Finnish news corpus for named entity recognition. *Language Resources and Evaluation* (2019), 1–26.

[17] Cicero Nogueira dos Santos and Victor Guimaraes. 2015. Boosting named entity recognition with neural character embeddings. *arXiv preprint arXiv:1505.05008* (2015).

[18] Burr Settles. 2004. Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*. 107–110.

[19] Vesa Siivola, Teemu Hirsimaki, Mathias Creutz, and Mikko Kurimo. 2003. Unlimited vocabulary speech recognition based on morphs discovered in an unsupervised manner. In *Eighth European Conference on Speech Communication and Technology*.

[20] Peter Smit, Sami Virpioja, Stig-Arne Grönroos, and Mikko Kurimo. 2014. Morfessor 2.0: Toolkit for statistical morphological segmentation. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*. 21–24.

[21] Peter Smit, Sami Virpioja, Mikko Kurimo, et al. 2017. Improved Subword Modeling for WFST-Based Speech Recognition.. In *INTERSPEECH*. 2551–2555.

[22] Cornelis Joost Van Rijsbergen. 1979. Information retrieval. (1979).

[23] Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A Smith, and Jaime Carbonell. 2018. Neural cross-lingual named entity recognition with minimal resources. *arXiv preprint arXiv:1808.09861* (2018).