Towards High Accuracy Named Entity Recognition for Icelandic

Svanhvít Ingólfsdóttir, Sigurjón Þorsteinsson, Hrafn Loftsson
Department of Computer Science
Reykjavik University
{svanhviti16, sigurjont, hrafn}@ru.is

Abstract

We report on work in progress which consists of annotating an Icelandic corpus for named entities (NEs) and using it for training a named entity recognizer based on a Bidirectional Long Short-Term Memory model. Currently, we have annotated 7,538 NEs appearing in the first 200,000 tokens of a 1 million token corpus, MIM-GOLD, originally developed for serving as a gold standard for part-of-speech tagging. Our best performing model, trained on this subset of MIM-GOLD, and enriched with external word embeddings, obtains an overall F_1 score of 81.3% when categorizing NEs into the following four categories: persons, locations, organizations and miscellaneous. Our preliminary results are promising, especially given the fact that 80% of MIM-GOLD has not yet been used for training.

1 Introduction

Named Entity Recognition (NER) is the task of identifying named entities (NEs) in text and labeling them by category. Before the work presented in this paper, no labeled data sets for NER existed for Icelandic. On the other hand, NER data sets exist for various other languages, e.g. for Spanish and Dutch (Tjong Kim Sang, 2002), for English and German (Tjong Kim Sang and De Meulder, 2003), and for seven Slavic languages (Piskorski et al., 2017). In all these data sets, NEs have been categorized into the following four categories: PER (person), LOC (location), ORG (organization), and MISC (miscellaneous), according to the CoNLL shared task conventions (Tjong Kim Sang, 2002).

The work in progress described in this paper is twofold. The first part consists of categorizing NEs in an Icelandic corpus, MIM-GOLD, containing about 1 million tokens, that has been developed to serve as a gold standard for training and evaluating part-of-speech (PoS) taggers (Loftsson et al., 2010). In the second part, MIM-GOLD is used to train and evaluate a named entity recognizer by applying a Bidirectional Long Short-Term Memory (BiLSTM) model (Hochreiter and Schmidhuber, 1997; Schuster and Paliwal, 1997). Our work will result in the first annotated Icelandic training corpus for NER and the first named entity recognizer for Icelandic based on machine learning (ML).

Currently, we have categorized 7,538 NEs appearing in the first 200,000 (200K) tokens of MIM-GOLD with the commonly used four NE categories: PER, LOC, ORG and MISC. Our best performing BiLSTM model, trained on this subset of MIM-GOLD, and enriched with external word embeddings (representations of words in n-dimensional space), obtains an overall F_1 score of 81.3%. Given the fact that 80% of MIM-GOLD has not yet been used for training, this preliminary result is promising and indicates that we may be able to develop a high accuracy named entity recognizer for Icelandic.

2 Background

In the last few years, neural network methods and deep learning have become the prevalent ML method in NER (Collobert et al., 2011; Yadav and Bethard, 2018). The main advantage of these methods is that they typically do not need domain-specific resources like lexicons or gazetteers (lists containing names of known entities) and features are normally inferred automatically as opposed to being learned with the help of hand-crafted feature templates as in feature-engineered systems.

Commonly used neural network architectures for NER include convolutional neural networks and recurrent neural networks (RNNs), along with

other ML methods, such as conditional random fields (Lafferty et al., 2001), which have been implemented as layers in neural network architectures (Lample et al., 2016) and used for NER tasks in under-resourced languages such as Persian (Poostchi et al., 2018).

Various studies show that pre-trained character and word embeddings are beneficial for NER tasks (Demir and Özgür, 2014; Wu et al., 2015; Dernoncourt et al., 2017). This is especially relevant for morphologically rich languages without large annotated datasets, such as Icelandic, since they offer a way to obtain subword information that cannot be inferred from the training corpus alone (Lafferty et al., 2001). Word embeddings have also been used to construct multilingual NER systems with minimal human supervision (Al-Rfou et al., 2014).

2.1 NeuroNER

Neural networks can be complicated and challenging to use, even for experts. *NeuroNER* (Dernoncourt et al., 2017) is an easy-to-use tool for NER based on a bidirectional RNN. An RNN is a neural network that is specialized for processing a sequence of values. A bidirectional RNN combines an RNN that moves forward in a sequence with another RNN that moves backward. The specific type of a bidirectional RNN used in NeuroNER is a BiLSTM model, which is capable of learning long-term dependencies.

The BiLSTM model in NeuroNER contains three layers: 1) a character-enhanced word-embedding layer, 2) a label prediction layer, and 3) a label sequence optimization layer (Dernoncourt et al., 2016). The first layer maps each token to a vector representation using two types of embeddings: a word embedding and a character-level token embedding. The resulting embeddings are then fed into the second layer which outputs the sequence of vectors containing the probability of each label for each corresponding token. Finally, the last layer outputs the most likely sequence of predicted labels based on the output from the previous label prediction layer.

Instead of implicitly learning the word embeddings, NeuroNER allows users to provide their own external (pre-trained) word embeddings (see Section 4).

NeuroNER enables users to annotate a corpus for NEs by interfacing with the web-based annotation tool *BRAT* (Stenetorp et al., 2012), and use the annotated corpus to train a named entity recognizer.

2.2 NER for Icelandic

As mentioned in Section 1, no labeled Icelandic data set for NER existed before our work started. Annotating a training corpus of a viable size for NER can be time-consuming task even if semi-automatic methods are used (Lample et al., 2016; Piskorski et al., 2017). Presumably, this is why no NER tools based on ML had been developed for Icelandic.

A rule-based named entity recognizer for Icelandic, *IceNER*, is part of the IceNLP toolkit (Loftsson and Rögnvaldsson, 2007). It has been reported to reach F_1 score of 71.5% without querying gazetteers, and 79.3% using a gazetteer (Tryggvason, 2009).

Greynir is an open-source NLP tool and website that parses sentences and extracts information from Icelandic news sites (Þorsteinsson et al., 2019). One of the features of Greynir is a rule-based named entity recognizer used to find and label person names in the news texts. The accuracy of this named entity recognizer has not been evaluated.

3 Developing the Training Corpus

The MIM-GOLD corpus is a balanced corpus of 1 million tokens of Icelandic texts, written in 2000-2010, from 13 different sources, including news texts, speeches from the Icelandic Parliament, laws and adjucations, student essays, and various web content such as blogs and texts from websites (Loftsson et al., 2010).[1] The texts have been tokenized and automatically PoS-tagged using the tagset developed for the Icelandic Frequency Dictionary corpus (Pind et al., 1991), with subsequent manual corrections (Helgadóttir et al., 2014). Note that MIM-GOLD is tagged for proper nouns, but does not contain any categorization of the proper nouns.

In order to reduce the work of categorizing Icelandic proper nouns in MIM-GOLD, we gathered official gazetteers of persons, organizations and place names, and used them as input to an automatic pre-classification program. Thereafter, we manually reviewed and corrected the results.

[1] Available for download from `http://malfong.is`

Category	*Count*	*%*
PER	3,045	40.4
LOC	1,748	23.2
ORG	1,768	23.4
MISC	977	13.0
Total	**7,538**	100.0

Table 1: Number of NEs in the 200K token training corpus.

Foreign tokens in the MIM-GOLD are all assigned the same tag with no further distinction, and since a large portion of them are NEs, they were reviewed and classified manually.

To make the review and correction process more efficient, we used the BRAT annotation tool (see Section 2). We use the IOB (inside, outside, beginning) format as used in the CoNLL data sets (Tjong Kim Sang, 2002).

At the time of writing, we have categorized 7,538 NEs appearing in the first 200K tokens of MIM-GOLD with the commonly used four categories: PER, LOC, ORG and MISC (see Table 1).

The annotated corpus was reviewed by a single linguist (first author), using the following definitions for each of the four categories:

- **Persons**: Names of humans and other beings, real or fictional, deities, pet names.
- **Locations**: Names of locations, real or fictional, i.e. buildings, street and place names, both real and fictional. All geographical and geopolitical entities such as cities, countries, counties and regions, as well as planet names and other outer space entities.
- **Organizations**: Icelandic and foreign companies and other organizations, public or private, real or fictional. Schools, churches, swimming pools, community centers, musical groups, other affiliations.
- **Miscellaneous**: All other capitalized nouns and noun phrases, such as works of art, products, events, printed materials, vessels and other named means of transportation, etc.

4 Training and Evaluation

The training corpus was arranged into two sets of different sizes, 100K and 200K tokens, each split into training (80%), validation (10%) and test (10%) sets. Four different models were trained and evaluated, for the two different training set sizes and for both implicitly and externally trained word embeddings.

W. embeddings	Implicit		External	
Corpus size	100K	200K	100K	200K
PER	71.8	76.1	95.2	93.3
LOC	61.8	65.6	81.8	85.6
ORG	23.5	40.5	62.7	69.2
MISC	3.2	28.3	14.8	41.5
Overall	**55.5**	**61.8**	**80.6**	**81.3**

Table 2: F_1 scores (%) of four different training configurations.

We pre-trained our own word embeddings of 200 dimensions using about 543 million tokens from a large unlabelled corpus, the Icelandic Gigaword Corpus (Steingrímsson et al., 2018), using a Word2Vec architecture (Mikolov et al., 2013).

All the parameters in NeuroNER's configuration file, controlling the structure of the model, along with the hyperparameters directed towards the learning process, were left at their default values. The only exception to this is the *token_embedding_dimension* parameter, controlling the length of the word vectors. This value was increased from 100 to 200 for the external word embeddings.

In the training, early stop was applied by default when no improvement had been seen on the validation set for ten consecutive epochs. The model used is based on the network weights taken from the epoch where the F_1 score last peaked for the validation set.

Evaluation was done automatically by NeuroNER according to CoNLL practices, which means that to score a true positive, both the NE category and the token boundaries need to be correct.

The F_1 scores for the models of the four training configurations, i.e. for the two training corpora sizes, with implicit and external word embeddings, are shown in Table 2. The best performing model is the one trained on 200K tokens and using external pre-trained word embeddings, achieving an overall F_1 score of 81.3%.

5 Discussion

The results presented in Section 4 are promising, especially given the few NEs found in the 200K tokens of the training corpus (see Table 1).

Table 2 shows that, using implicitly trained word embeddings, the F_1 score increases considerably when doubling the corpus size, i.e. from 55.5% to 61.8%. This was to be expected, as the training set in the 100K token corpus only contains around 80K tokens, and has thus a very limited number of NE examples to learn from. When further increasing the training corpus, we expect this trend to continue.

However, a more effective approach to increase the accuracy proved to be incorporating pre-trained word embeddings. In that case, the F_1 score increases to 81.3% when using the 200K corpus. In what follows, we refer to this best performing model as *200K_External*.

Most studies on the benefits of word embeddings in NER tasks do not report more than a few percent points increase in F_1 score by introducing pre-trained word embeddings. Intuitively, we deduce that the main reason we are experiencing this huge benefit of pre-trained word embeddings is the small size of our training corpus. For a small training set, the model will more often encounter unseen words in the test set. In our 200K corpus, 60% of the incorrectly labeled words had not been seen in the training set. When the large collection of word embeddings is added to the pool, the chances that a word is known increase substantially.

Another reason as to why word embeddings from a large external corpus are so beneficial for our model may be the underlying language. Icelandic, a morphologically rich language, presents special challenges for various NLP tasks, such as NER. Nouns, generally the building blocks of NEs, have up to 16 unique inflectional forms, and verbs and adjectives can have over a hundred different forms. This greatly increases the vocabulary size of a corpus, and causes a problem with data sparsity, as pointed out by Demir and Özgür (2014), for the case of Turkish and Czech. The implication is that a NER system may not recognize a NE in the test set even if it has seen it in a different form in the training set.[2] We could try to lemmatize the tokens in the training corpus and use the normalized output for building the NER model, but then we would lose important contextual information about the NEs and their neighbors. The pre-trained word embeddings contribute many examples to the model of different word forms that do not appear in the training set, and the likelihood of correctly labeling them increases as a result.

With a larger corpus, say by doubling it once again, we believe, from the trend, that the F_1 score without external embeddings might end up between the earlier score (61.8%) and the one obtained by 200K_External (81.3%). On the other hand, the results with pre-trained embeddings indicate a much slower increase in F_1 score when increasing the size of the corpus (from 80.6% to 81.3% when increasing the corpus size from 100K to 200K). This might indicate that we are approaching the upper limit with regard to the F_1 score.

NeuroNER has achieved 90.5% F_1 score for English on the CoNLL data set (Dernoncourt et al., 2017). This English data set contains 35,089 NEs (Tjong Kim Sang and De Meulder, 2003) whereas our 200K Icelandic training corpus contains only 7,538 NEs. Therefore, we are optimistic that increasing the training corpus size for Icelandic will further increase the overall F_1 score, albeit we do not expect getting close to the score for English, which is a morphologically simple language compared to Icelandic.

5.1 Accuracy for Different Categories

Table 2 shows a considerable difference in the accuracy of different categories. Especially promising are the results for the PER category, with F_1 score of 93.3% for 200K_External. The recall for PER is high, 94.85%, which means that only about 5% of the person names in the test set were not identified. Several factors may explain the top performance in this category. Most importantly, person names are by far the most frequent entity type in the training corpus, almost double that of the LOC and ORG categories (see Table 1). Person names are often constructed in a similar manner, with Icelandic full names usually composed of one or two given names and a surname ending in *-son* or *-dóttir* "daughter". Furthermore, they are almost always capitalized, and since they are not unique (many people can have the same name), each person name is bound to appear more often than, for example, each organization name.

200K_External also performs quite well on the LOC category (85.6%). It is the nature of a corpus sampled from any geographic area that some lo-

[2]For example, the Icelandic person name "Egill" (nominative) may be tagged as such in the training set and then appear as "Egil" (accusative),"Agli" (dative), or "Egils" (genitive) in the test set.

	Predicted categories				
	LOC	MISC	ORG	PER	O
LOC	161	3	11	8	5
MISC	7	87	26	14	68
ORG	19	6	174	4	21
PER	0	3	1	568	16
O	1	46	33	12	19,245

Table 3: Confusion matrix for the classification of the test set in 200K_External. True categories are shown vertically, predicted categories horizontally. The *O* category denotes "outside", i.e. that the corresponding token is not a part of a NE.

cations appear more often than others, in this case *Ísland* "Iceland" and *Reykjavík*, to name two of the most common. This means that during testing, the system is much more likely to label them correctly, because they are likely to have been found during training. Another property of place names is that they tend to be single word entities, and are capitalized, with a few exceptions. As a result, detecting word boundaries becomes less of a challenge.

The LOC and ORG categories are equally common in the corpus, but the accuracy for ORG (69.2%) is significantly worse. In the ORG category, word boundaries are a problem, as organizations are often composed of more than one word, not necessarily capitalized, e.g. *Samband lífeyrisþega ríkis og bæja* "Organization (of) pensioners (of) state and towns". Furthermore, sometimes it can be hard to decide whether an entity is an organization name or a product, which may cause overlap with the MISC category.

The MISC category was the most problematic one for 200K_External, with F_1 score of only 41.5%. Recall is particularly low (33.6%), meaning many MISC entities are not found, and precision is not particularly high either (54.17%), thus many entities are mislabeled.

The confusion matrix (see Table 3) from the classification of the test set in 200K_External shows how many of the tokens in the test set were correctly labeled (the diagonal line running from the top left corner to the bottom right), and where mislabeling occurs. The MISC category contains the most outliers, with a total of 115 NEs mislabeled out of 202 in total. In the PER category only 20 NEs out of 588 are mislabeled. There are only around 1000 MISC entities in the whole 200K corpus, and a lot of variation in how they are constructed, which makes detecting them harder, even for human annotators. Some are long book or movie titles, some are complicated product names with numbers and hyphens, and there is no correlation within the category. This is the category that tends to score lowest in most NER models, but a substantially larger corpus should lead to some improvement.

6 Conclusion

We have described work in progress consisting of annotating the MIM-GOLD corpus for NEs and using it to train a named entity recognizer based on a BiLSTM model. By only categorizing about 20% of the NEs found in MIM-GOLD, the best resulting model, enriched with external word embeddings, achieves an overall F_1 score of 81.3%. We are optimistic that we can further increase the F_1 score for Icelandic by increasing the training corpus size. Currently, the number of NEs found in our training corpus (7,538) is only about $1/5$ of the training examples provided in the English CoNLL data set.

In future work, we will continue categorizing NEs in MIM-GOLD, such that we will be able to use 100% of the corpus to train NER models for Icelandic. We are also working on adding categories for numerical units, such as dates and prices. The annotated corpus will be publicly released, in order to serve as a valuable asset for further research on NER for Icelandic. The resulting NER models will also be made available for public use.

In addition to further developing our BiLSTM model and testing different configurations, we intend to develop models based on other ML techniques, for the sake of comparison, as well as being able to combine various different classifiers.

As mentioned in Section 2, different word and character representations have shown promise when developing NER models. In this preliminary work we used the Word2Vec architecture, which resulted in a large improvement, but in the future we intend to measure how some of the other word and character representation methods compare, e.g. contextual word embeddings such as ELMo (Peters et al., 2018) and flair (Akbik et al., 2018).

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual String Embeddings for Sequence Labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, COLING 2018, Santa Fe, New Mexico, USA.

Rami Al-Rfou, Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. 2014. POLYGLOT-NER: Massive Multilingual Named Entity Recognition. *CoRR*, abs/1410.3791.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *CoRR*, abs/1103.0398.

Hakan Demir and Arzucan Özgür. 2014. Improving Named Entity Recognition for Morphologically Rich Languages Using Word Embeddings. In *13th International Conference on Machine Learning and Applications*, Detroit, MI, USA.

Franck Dernoncourt, Ji Young Lee, and Peter Szolovits. 2017. NeuroNER: an easy-to-use program for named-entity recognition based on neural networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Copenhagen, Denmark.

Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. 2016. De-identification of patient notes with recurrent neural networks. *CoRR*, abs/1606.03475.

Sigrún Helgadóttir, Hrafn Loftsson, and Eiríkur Rögnvaldsson. 2014. Correcting Errors in a New Gold Standard for Tagging Icelandic Text. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, LREC 2014, Reykjavik, Iceland.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural computation*, 9(8):1735–1780.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML 2001, Williamstown, MA, USA.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural Architectures for Named Entity Recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, California, USA.

Hrafn Loftsson and Eiríkur Rögnvaldsson. 2007. IceNLP: A Natural Language Processing Toolkit for Icelandic. In *Proceedings of InterSpeech, Special Session: Speech and language technology for less-resourced languages*, Antwerp, Belgium.

Hrafn Loftsson, Jökull H. Yngvason, Sigrún Helgadóttir, and Eiríkur Rögnvaldsson. 2010. Developing a PoS-tagged corpus using existing tools. In *Proceedings of 7th SaLTMiL Workshop on Creation and Use of Basic Lexical Resources for Less-Resourced Languages*, LREC 2010, Valetta, Malta.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *CoRR*, abs/1301.3781.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana, USA.

Jörgen Pind, Friðrik Magnússon, and Stefán Briem. 1991. *Íslensk orðtíðnibók [The Icelandic Frequency Dictionary]*. The Institute of Lexicography, University of Iceland, Reykjavik, Iceland.

Jakub Piskorski, Lidia Pivovarova, Jan Šnajder, Josef Steinberger, and Roman Yangarber. 2017. The First Cross-Lingual Challenge on Recognition, Normalization, and Matching of Named Entities in Slavic Languages. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, Valencia, Spain.

Hanieh Poostchi, Ehsan Zare Borzeshi, and Massimo Piccardi. 2018. BiLSTM-CRF for Persian Named-Entity Recognition ArmanPersoNERCorpus: the First Entity-Annotated Persian Dataset. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, LREC 2018, Miyazaki, Japan.

Mike Schuster and Kuldip. K. Paliwal. 1997. Bidirectional Recurrent Neural Networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Steinþór Steingrímsson, Sigrún Helgadóttir, Eiríkur Rögnvaldsson, Starkaður Barkarson, and Jón Guðnason. 2018. Risamálheild: A Very Large Icelandic Text Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, LREC 2018, Miyazaki, Japan.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, Avignon, France.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 Shared Task: Language-Independent

Named Entity Recognition. In *Proceedings of CoNLL-2002*, Taipei, Taiwan.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of CoNLL-2003*, Edmonton, Canada.

Aðalsteinn Tryggvason. 2009. Named Entity Recognition for Icelandic. Research report – Reykjavik University.

Yonghui Wu, Jun Xu, Min Jiang, Yaoyun Zhang, and Hua Xu. 2015. A Study of Neural Word Embeddings for Named Entity Recognition in Clinical Text. *AMIA Annual Symposium Proceedings*, 2015:1326–33.

Vikas Yadav and Steven Bethard. 2018. A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*, COLING 2018, Santa Fe, New Mexico, USA.

Vilhjálmur Þorsteinsson, Hulda Óladóttir, and Hrafn Loftsson. 2019. A Wide-Coverage Context-Free Grammar for Icelandic and an Accompanying Parsing System. In *Proceedings of Recent Advances in Natural Language Processing (to appear)*, RANLP 2019, Varna, Bulgaria.