Named Entity Recognition for Icelandic: Annotated Corpus and Models

Svanhvít L. Ingólfsdóttir[0000−0001−6965−692X],
Ásmundur A. Guðjónsson[0000−0002−7790−664X], and
Hrafn Loftsson[0000−0002−9298−4830]

Language and Voice Lab
Reykjavik University
Menntavegur 1, 101 Reykjavík, Iceland
svanhviti16@ru.is,asmundur10@ru.is,hrafn@ru.is
https://lvl.ru.is

Abstract. Named entity recognition (NER) can be a challenging task, especially in highly inflected languages where each entity can have many different surface forms. We have created the first NER corpus for Icelandic by annotating 48,371 named entities (NEs) using eight NE types, in a text corpus of 1 million tokens. Furthermore, we have used the corpus to train three machine learning models: first, a CRF model that makes use of shallow word features and a gazetteer function; second, a perceptron model with shallow word features and externally trained word clusters; and third, a BiLSTM model with external word embeddings. Finally, we applied simple voting to combine the model outputs. The voting method obtains an F_1 score of 85.79, gaining 1.89 percentage points compared to the best performing individual model. The corpus and the models are publicly available.

Keywords: Named entity recognition · Corpus annotation · BiLSTM · CRF · Clustering · Machine learning

1 Introduction

Since the integration of named entity recognition (NER) into the sixth Message Understanding Conference (MUC) in 1995 [16], NER has become recognized as an important task in natural language processing (NLP), and NER datasets and methods have been developed for many languages. Detecting and recognizing named entities (NEs) in text is not a trivial task, as various patterns need to be learned, and new proper names appear frequently. Furthermore, in highly inflected languages, such as Icelandic, each proper name can have many different surface forms, which further complicates the task.

No NER corpus was available for the Icelandic language before the work introduced in [18], where a sample of 200,000 tokens, from the MIM-GOLD corpus of 1 million tokens [23], was annotated with four NE types and used for training a NER prototype. In this paper, we describe the completion of the annotation of the whole corpus, using eight NE types, which has resulted in 48,371 NEs. The corpus, MIM-GOLD-NER, is available online[1]. Furthermore, we describe the evaluation of three different models trained on the corpus: first, a Conditional Random Field (CRF) model that makes use of shallow word features and a gazetteer function; second, a perceptron model with shallow word features and externally trained word clusters; and third, a bidirectional long short-term memory (BiLSTM) model with external word embeddings. Finally, we have combined our model outputs using a simple voting method, obtaining an F_1 score of 85.79 and gaining 1.89 percentage points compared to the best performing individual model. The code for our models is available online[2].

2 Related Work

The most commonly annotated entity types in NER corpora in the general domain are PERSON, LOCATION and ORGANIZATION, both due to the fact that these entities are very common in general texts and newswire texts, which are often the source of NER corpora, and that they contain information that may be valuable for a variety of purposes. The three generic NE types were first proposed for the MUC-6 event as a subtask called ENAMEX [16]. The other subtasks in MUC-6 are named TIMEX (dates and times) and NUMEX (monetary values and percentages). In the CoNLL shared task, a fourth category, MISCELLANEOUS was introduced, which covers proper names that fall outside of the three classic categories, PERSON, LOCATION and ORGANIZATION [31].

While rule-based methods for NER can be highly efficient, especially in a well-defined domain [10], machine learning (ML) methods took over from rule-based method as the main approach to NER as the field developed and data became more readily available. These range from unsupervised to fully supervised methods (e.g. [3, 25, 35]), as well as hybrid systems, combining various supervised and unsupervised methods and carefully engineered features, e.g. [2].

Although the majority of the early work on NER was conducted on the English language, NER corpora and published work is now available for various different languages. This is an important development, because of the many structural, morphological, and orthographic features that characterize different languages. Enriching the input representation with pre-trained word embeddings has, for example, proven useful for NER in languages such as Turkish [12] and Arabic [21], as have the larger and more complex language models that have recently become popular, such as the Finnish [33] and Slavic [4] BERT models.

Various published work exists for the languages most related to Icelandic, i.e. the Scandinavian languages. The most recent work for Swedish includes a BiLSTM model [34]. A Danish NER corpus is presented in [13], but some of the latest research for Danish NER focuses on cross-lingual transfer, to make up for the limited data available [27]. For Norwegian, the most recent work includes [19], which involves new annotated NER datasets for the two written forms of

[1] http://hdl.handle.net/20.500.12537/42

[2] http://github.com/cadia-lvl/NER

Norwegian, Bokmål and Nynorsk, and BiLSTM models enriched with pre-trained word embeddings.

Limited work exists on NER for the Icelandic language, most likely due to the lack of an annotated NER corpus. IceNER, a rule-based NER system is part of the IceNLP toolkit [22]. It has been reported to reach an F_1 score of 71.5 without querying a gazetteer, and 79.3 with a gazetteer lookup [32].

3 Corpus Annotation

As a basis for our NER corpus, we used MIM-GOLD, the Icelandic Gold Standard corpus [23], a balanced text corpus of approximately 1 million tokens, tagged with part-of-speech (PoS). All the texts are from the years 2000–2009 and are sourced from thirteen text types, including news texts, books, blogs, websites, laws, adjudications, school essays, scripted radio news, web media texts and emails. For most languages, this variety of text genres is not common in NER corpora, which are often centered on newswire texts (with some exceptions, such as the Portuguese HAREM NER contests [29, 28]). Nevertheless, our corpus, MIM-GOLD-NER, is heavy on news-related content, as newspaper articles, web media, and radio news account for 36% of the tokens.

3.1 Annotation Process

Eight NE types are tagged in MIM-GOLD-NER: PERSON, LOCATION, ORGANIZATION, MISCELLANEOUS, DATE, TIME, MONEY and PERCENTAGE. The first four entity types are the same as used in the CoNLL shared tasks. The last four NE types were adapted from the NUMEX and TIMEX types in the MUC events.

We applied a semiautomatic approach when annotating the corpus, using gazetteers and regular expressions to extract as many entities as possible before reviewing and correcting the corpus manually.

Preprocessing Gazetteers were collected from official Icelandic resources. For extracting the person names, we used the Database of Modern Icelandic Inflections (DMII) [6]. Additionally, we collected lists of place names and addresses, as well as company names. In the end, we had gazetteers with 15,000 person names, 97,000 location names, and 90,000 organization names. Since the original MIM-GOLD is PoS-tagged, we were able to use the PoS tags to filter out likely proper nouns and match them with our gazetteers, to produce NE candidates. Heuristics and knowledge of the language were used to resolve doubts and ambiguities and to try to determine NE boundaries. Remaining ambiguities were registered to be resolved manually. Regular expressions were used for automatically extracting the numerical entities, taking care to match entities regardless of how they are written out in the corpus, whether numerically or alphabetically.

Manual review After the automatic preprocessing step, the next task was reviewing the resulting annotations, fixing errors and picking up any remaining NEs missed in the previous step, since some inaccuracies were bound to appear, both in the classification and the span (boundaries) of the entities.

Guidelines were constructed regarding the taxonomy used for the annotation of the corpus. For the four NE types adapted from CoNLL, we mostly relied on the CoNLL guidelines for each entity type [9], though with some modifications to fit Icelandic settings and writing conventions. The four numerical and temporal entity types, DATE, TIME, MONEY, and PERCENT, appeared less commonly in the corpus, and were easier to find using regular expressions. The MUC guidelines were followed as closely as possible when annotating these entities.

One implication of using a balanced corpus such as MIM-GOLD, is that only parts of it have been reviewed/edited, so the texts vary quite a lot in writing quality; some have been thoroughly proofread (published books, laws and adjudications), some have undergone some editing (news articles, some web content, scripted texts for radio), and some have not been edited at all (blogs, emails, web content, classified newspaper ads). The corpus contained many problematic NE candidates, for which annotation was not clear. These were marked specially during the annotation and resolved at the end, to keep consistency within the corpus. In some parts of the corpus, such as laws and regulatory texts, NEs are sparsely distributed, but we have nonetheless opted for annotating the whole corpus without removing sentences unlikely to contain NEs. This was done both to keep the organic distribution of NEs in the corpus and to ensure compatibility with the original MIM-GOLD corpus.

One annotator (the first author of this paper) was in charge of defining the annotation task and labeling the bulk of the corpus. This was done part-time in the course of a year. A second annotator (the second author) stepped in to help in the last weeks, and reviewed around 8% of the corpus. We estimate that 150-200 hours went into the manual annotation stage. The annotation resulted in 48,371 NEs, split between the eight entity types as shown in Table 1.

Table 1. NE split in the MIM-GOLD-NER corpus.

Entity type	Total NE count	Percentage
Person	15,599	32.25%
Location	9,011	18.63%
Organization	8,966	18.54%
Miscellaneous	6,264	12.95%
Date	5,529	11.43%
Time	1,214	2.51%
Money	1,050	2.17%
Percent	738	1.53%
Total	48,371	

Once the annotation was finished, an external linguist reviewed a randomly chosen 10% sample of the corpus, to estimate the accuracy of the annotation. The corpus sample contained a total of 4,527 NEs. The reviewer was presented with the same instructions and guidelines as annotator 1 and annotator 2, and was asked to mark any doubts or errors spotted in the corpus sample. Annotator 1 then reviewed this list of error candidates and, in accordance with the guidelines, evaluated which of them were true positives, i.e. real errors that should be fixed. In the end, 250 error candidates were found, out of which 180 were marked as false positives by annotator 1. These false positives were due to either doubts that the reviewer had on how to annotate or lack of detail in the annotation guidelines. Thus, the total number of real errors was 70, which is equivalent to an accuracy of 98.45% for this corpus sample.

4 Models

Until recently, the main approach for NER has been the application of BiLSTM models, along with the best hybrid feature-engineered ML systems [36]. Even though these methods have mostly fallen in the shadow of state-of-the-art transformer models [5, 15], they do a good job of solving problems where the input is sequential data, such as in NER. Since a transformer model does not yet exist for Icelandic, we opted for these tried and tested methods for the first experiments with our new corpus.

Three different methods were chosen for the experiments: a CRF model, a perceptron model, and a BiLSTM model. The three models were then combined into one ensemble NER tagger using simple voting. Before presenting our results, we will briefly describe each method.

4.1 Conditional Random Field (CRF)

CRF is a conditional probabilistic modeling method, which can take context into account [20]. Passos et al. [25] implemented a stacked linear-chain CRF system for NER that makes use of shallow word features in their baseline model, along with gazetteers, and then compared the results when adding Word2Vec embeddings [24] and Brown clusters [8], among other things. Their best performing model achieved an F_1 score of 90.9 on the English CoNLL-03 test set.

We implemented a model inspired by this baseline system. Our model uses the following word features:

- the word lower-cased
- word suffixes, length 1 to 4
- a boolean for whether the word is all in uppercase letters
- a boolean for whether the first letter of the word is in uppercase
- a boolean for whether the word is a digit
- all the character n-grams within the word, of length 2 to 5
- the four words prior to the word

- the four words after the word

These parameters were mostly found by trial and error. If a parameter did not improve the model, or if removing a parameter made no difference on the validation set, it was discarded. As an example, suffixes are used, but not prefixes, since the prefixes did not contribute any difference. The gazetteers used for the corpus annotation stage were reused for this model, and each list had its own boolean parameter for whether the word appeared in the list.

4.2 IXA-pipe-nerc

In a survey on different NER architectures [36], the best performing non-neural model was *IXA-pipe-nerc*, a NER module which is a part of the IXA pipes NLP tool [1]. IXA-pipe-nerc is based on a perceptron model and utilizes both shallow local word features and semi-supervised word clusters [2]. It supports including PoS tags as features, as well as a gazetteer lookup. The clustering features used were Brown [8], Clark [11], and Word2Vec [24] clusters. The software is open source and language-independent, making it straightforward to train a model for a new language. The supported local word features are:

- current lower-cased token
- token shape
- the previous prediction for the word
- whether it is the first token in the sentence
- both the prefixes and suffixes of the token, with default of length 4
- word bigrams and trigrams, which both include the current token and the token shape
- all the character n-grams within the word, with default of length 2 to 5

Evaluation shows that the biggest performance boost comes from the word clusters. The best model in [2] achieved an F_1 score of 91.36 on the English CoNLL-03 test set.

The configuration for our model was chosen by experimenting on the validation set, selecting the features that gave the best results. In our model, we used the default values for the features enumerated above, except we used character n-grams of length 1 up to 11, as well as the word trigrams, which are disabled by default. We trained the three types of word clusters on the Icelandic Gigaword Corpus (IGC) [30], a corpus of around 1.4 billion words of Icelandic texts from various sources.

4.3 BiLSTM with Pre-trained Word Embeddings

For the BiLSTM experiments, we used a program called *NeuroNER* [14]. NeuroNER is described as an easy-to-use program for training models to recognize and classify NEs. It uses TensorFlow[3] for training the neural networks and has

[3] `https://www.tensorflow.org/`

been reported to reach an F_1 of 90.5 on the English CoNLL-03 dataset. NeuroNER is divided into three layers. First, a BiLSTM 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 fed into the second layer, a BiLSTM which outputs the sequence of vectors containing the probability of each label for each corresponding token. Finally, a CRF decoding 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 offers the possibility to incorporate external word embeddings, pre-trained on a larger dataset. We have provided external word embeddings, trained on the 2018 version of the IGC. For the sake of comparison, Word2Vec [24], GloVe [26], as well as FastText [7] embeddings were tested. GloVe embeddings turned out to give slightly better results than the other two, so they were used in the models presented in our results. For the GloVe embeddings, dimensions were set at 300, window size at 10, and the minimum term count at 5. NeuroNER was configured with the following main parameters:

- character embedding dimension = 25
- character lstm dimension = 25
- dropout = 0.5
- patience = 10
- maximum number of epochs = 100
- optimizer = stochastic gradient descent
- learning rate = 0.005

4.4 CombiTagger

Different ML models may have different strengths and weaknesses depending on the methods used to train them. For NER, some may, for example, work better on the more regular numerical entities, while others may be better at overcoming misspellings in the text. One way of leveraging the strengths of different models, for an increased overall performance, is using a voting system for the output tags. CombiTagger [17], a system originally developed for combining different PoS taggers, offers simple and weighted voting. In our work, we fed our three best NER model outputs into CombiTagger and applied simple voting.

5 Results and Discussion

In this section, we present and discuss the results from training on MIM-GOLD-NER using the methods described in Section 4.

For the total corpus of 1 million tokens, the split between training, validation and test sets was 80%, 10% and 10%, respectively, which gave a training set of around 800,000 tokens, and validation and test sets of around 100,000 tokens each. We also trained on five different sizes of the corpus, keeping the test set intact in all experiments to maintain consistency in the evaluation. The CoNLL-03 evaluation metrics are used, meaning that both the type and the boundaries of a predicted NE need to match the gold label for it to count as correct.

Table 2. F_1 for the different models, in addition to CombiTagger, which combines the output from the three best models.

	Overall	PER	LOC	ORG	MISC	DATE	TIME	MON	PERC
BiLSTM-internal	73.60	80.11	74.98	65.82	44.10	86.73	91.83	81.86	92.73
CRF	82.24	87.18	86.04	77.02	53.87	90.90	94.46	84.30	98.21
IXA-pipe-nerc	83.10	87.90	85.54	77.02	61.57	91.96	94.29	85.59	97.35
BiLSTM-GloVE	83.90	89.53	85.45	79.03	61.77	90.60	94.78	89.45	95.54
CombiTagger	**85.79**	90.19	88.21	81.23	64.27	93.13	96.41	86.58	98.65

In Table 2, the results from training the different models on the whole corpus are presented. Before considering the best performing method, CombiTagger, we will discuss how the CRF, IXA-pipe-nerc, and BiLSTM models performed overall, and on each NE type.

It is not surprising that the most advanced model, BiLSTM, outperforms the other two in overall score, but it is interesting to see how close behind the IXA-pipe-nerc model comes – by only 0.80 percentage points. Furthermore, IXA-pipe-nerc, which uses externally trained (semi-supervised) word clusters, only improves from the CRF by 0.86 percentage points. In [36], the IXA-pipe-nerc model also obtained marginally better scores (0.46 percentage points) than a CRF model, when evaluated on English.

Looking at the individual entity types, we see that the BiLSTM outperforms the other two in five out of eight types, which should not come as a surprise. What was more unexpected, however, is that the CRF, despite scoring somewhat worse overall, outperforms the IXA-pipe-nerc model in two categories, LOCATION and PERCENT. The reason for the high-scoring LOCATION type may be the gazetteer lookup, implemented as part of the CRF model, since the place names gazetteer was quite exhaustive. The IXA-pipe-nerc model, however, gives the best results for DATE, outperforming the others by over 1 percentage point.

Note the effect of using pre-trained word embeddings as external input into the BiLSTM. BiLSTM-internal was trained without these external word embeddings, i.e. word and character embeddings were trained internally using the training data itself. In contrast, BiLSTM-GloVe uses externally trained GloVe embeddings. We attribute this gain from using the pre-trained word vectors to the fact that Icelandic, being a highly inflected language, has so many surface forms for any lemma, that even though during training the network has seen one surface form of a word, it doesn't know the next time it sees a different surface form that it is the same word. Incorporating a pool of word vectors trained from a corpus of 1.4 billion tokens gives the network access to information on many different word forms.

The CombiTagger ensemble method improves the overall results, with an F_1 score of 85.79, which is better than any of the individual models. When it is compared with our best NER model, we see a 1.89 percentage points improvement in overall F_1, and 2.69 points when compared with the second best model. The observed results show that while the three different models are not equal in quality, each one of them is better than the other two at predicting some particular NE type. The different models thus tend to produce different (complementary) errors and the differences can be exploited to yield better results. Therefore, CombiTagger outperforms all the others on all but one NE types.

Table 3 shows that training on increasing sizes of the data gradually improves the performance. However, even with a dataset of 540,000 tokens, the overall F_1 is 82.37, and 85.17 in the 720,000 token dataset, which is not far behind from the result obtained on the whole corpus. Furthermore, preliminary experiments with training on a subset of the text types, containing only news texts, indicate that this may be a viable approach, especially if the intended use is within a particular domain.

Table 3. F_1 scores for CombiTagger on different sizes of the data. For clarification of the model names, CombiTagger-180K stands for a corpus size of around 180,000 tokens, with a training data size of around 160,000 tokens, validation set size of approximately 20,000 tokens, and the consistent test set size of 100,693 tokens (10% of each corpus size was reserved for the test set).

	Overall	PER	LOC	ORG	MISC	DATE	TIME	MON	PERC
CombiTagger-180K	76.67	85.50	81.87	69.18	46.34	79.60	89.75	71.50	96.46
CombiTagger-360K	79.14	88.24	82.13	71.59	54.11	84.34	86.89	78.10	97.78
CombiTagger-540K	82.37	89.23	83.75	76.00	58.92	89.77	91.97	87.67	98.20
CombiTagger-720K	85.16	89.19	87.19	81.96	61.16	92.59	95.40	90.35	97.78
CombiTagger-900K	85.79	90.19	88.21	81.23	64.27	93.13	96.41	86.58	98.65

6 Conclusion

We have described the annotation of the first NER corpus for Icelandic and the initial experiments on using the data for training and evaluating Icelandic NER models. This corpus, with 48,371 NEs tagged in 1 million tokens, is one of the largest manually annotated NER corpora we have come across in the literature, and includes a variety of text types that have been annotated for eight common entity types.

Several different model architectures and training set sizes were tested on the data, and an ensemble method using simple voting from three models was shown to perform considerably better than any individual model. These results are presented without any post-processing, such as a gazetteer lookup, commonly used to boost NER results. The morphological intricacies of Icelandic make NER a nuanced problem, but based on these first results we are optimistic about obtaining higher scores with more advanced models in the future, e.g. by using BERT-type models.

Acknowledgments

This work was funded by the Icelandic Strategic Research and Development Programme for Language Technology 2019, grant no. 180027-5301.

References

1. Agerri, R., Bermudez, J., Rigau, G.: IXA pipeline: Efficient and Ready to Use Multilingual NLP tools. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation. LREC 2014, Reykjavik, Iceland (2014)
2. Agerri, R., Rigau, G.: Robust Multilingual Named Entity Recognition with Shallow Semi-Supervised Features. arXiv e-prints arXiv:1701.09123 (2017)
3. Ahmed, I., Sathyaraj, R.: Named Entity Recognition by Using Maximum Entropy. International Journal of Database Application and Theory **8**, 43–50 (2015), https://doi.org/10.14257/ijdta.2015.8.2.05
4. Arkhipov, M., Trofimova, M., Kuratov, Y., Sorokin, A.: Tuning Multilingual Transformers for Language-Specific Named Entity Recognition. In: Proceedings of the 7^{th} Workshop on Balto-Slavic Natural Language Processing. Florence, Italy (2019), https://doi.org/10.18653/v1/W19-3712
5. Baevski, A., Edunov, S., Liu, Y., Zettlemoyer, L., Auli, M.: Cloze-driven Pretraining of Self-attention Networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9^{nd} International Joint Conference on Natural Language Processing. EMNLP/IJCNLP, Hong Kong, China (2019), https://doi.org/10.18653/v1/D19-1539
6. Bjarnadóttir, K.: The Database of Modern Icelandic Inflection. In: Proceedings of the "Language Technology for Normalisation of Less-Resourced Languages" (SaLTMiL 8 – AfLaT2012), workshop at the 8^{th} International Conference on Language Resources and Evaluation. LREC 2012, Istanbul, Turkey (2012)
7. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics **5**, 135–146 (2017), https://doi.org/10.1162/tacl_a_00051
8. Brown, P.F., deSouza, P.V., Mercer, R.L., Pietra, V.J.D., Lai, J.C.: Class-Based n-Gram Models of Natural Language. Computational Linguistics **18**(4), 467–479 (1992), https://www.aclweb.org/anthology/J92-4003
9. Chinchor, N., Brown, E., Ferro, L., Robinson, P.: Named Entity Recognition Task Definition. Tech. Rep. Version 1.4, The MITRE Corporation and SAIC (1999)
10. Chiticariu, L., Krishnamurthy, R., Li, Y., Reiss, F., Vaithyanathan, S.: Domain Adaptation of Rule-Based Annotators for Named-Entity Recognition Tasks. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. EMNLP 2010, Cambridge, MA, USA (2010), http://aclweb.org/anthology/D10-1098
11. Clark, A.: Combining Distributional and Morphological Information for Part of Speech Induction. In: Proceedings of the 10^{th} Conference of the European Chapter of the Association for Computational Linguistics. EACL 2003, Budapest, Hungary (2003), https://www.aclweb.org/anthology/E03-1009

12. Demir, H., Özgür, A.: Improving Named Entity Recognition for Morphologically Rich Languages Using Word Embeddings. In: Proceedings of the 13^{th} International Conference on Machine Learning and Applications. ICMLA 2013, Miami, FL, USA (2014), https://doi.org/10.1109/ICMLA.2014.24
13. Derczynski, L., Field, C.V., Bøgh, K.S.: DKIE: Open Source Information Extraction for Danish. In: Proceedings of the Demonstrations at the 14^{th} Conference of the European Chapter of the Association for Computational Linguistics. EACL 2014, Gothenburg, Sweden (2014), https://doi.org/10.3115/v1/E14-2016
14. Dernoncourt, F., Lee, J.Y., Szolovits, P.: NeuroNER: an easy-to-use program for named-entity recognition based on neural networks. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. EMNLP 2017, Copenhagen, Denmark (2017)
15. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). NAACL, Minneapolis, MN, USA (2019), https://doi.org/10.18653/v1/N19-1423
16. Grishman, R., Sundheim, B.: Message Understanding Conference-6: A Brief History. In: Proceedings of the 16^{th} Conference on Computational Linguistics – Volume 1. COLING 1996, Copenhagen, Denmark (1996), https://www.aclweb.org/anthology/C96-1079/
17. Henrich, V., Reuter, T., Loftsson, H.: CombiTagger: A System for Developing Combined Taggers. In: Proceedings of the 22^{nd} International FLAIRS Conference, Special Track: "Applied Natural Language Processing". Sanibel Island, FL, USA (2009), https://www.aaai.org/ocs/index.php/FLAIRS/2009/paper/viewFile/67/296
18. Ingólfsdóttir, S.L., Þorsteinsson, S., Loftsson, H.: Towards High Accuracy Named Entity Recognition for Icelandic. In: Proceedings of the 22^{nd} Nordic Conference on Computational Linguistics. NoDaLiDa 2019, Turku, Finland (2019), https://www.aclweb.org/anthology/W19-6142
19. Johansen, B.: Named-Entity Recognition for Norwegian. In: Proceedings of the 22^{nd} Nordic Conference on Computational Linguistics. NoDaLiDa 2019, Turku, Finland (2019), https://www.aclweb.org/anthology/W19-6123
20. Lafferty, J.D., McCallum, A.K., Pereira, F.C.N.: 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 (2001)
21. Liu, L., Shang, J., Han, J.: Arabic Named Entity Recognition: What Works and What's Next. In: Proceedings of the Fourth Arabic Natural Language Processing Workshop. Florence, Italy (2019), https://doi.org/10.18653/v1/W19-4607
22. Loftsson, H., Rögnvaldsson, E.: IceNLP: A natural language processing toolkit for Icelandic. In: Proceedings of the Annual Conference of the International Speech Communication Association. Antwerp, Belgium (2007)
23. Loftsson, H., Yngvason, J.H., Helgadóttir, S., Rögnvaldsson, E.: Developing a PoS-tagged corpus using existing tools. In: Proceedings of "Creation and use of basic lexical resources for less-resourced languages", workshop at the 7^{th} International Conference on Language Resources and Evaluation. LREC 2010, Valetta, Malta (2010)
24. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space. arXiv e-prints arXiv:1301.3781 (2013)

25. Passos, A., Kumar, V., McCallum, A.: Lexicon infused phrase embeddings for named entity resolution. In: Proceedings of the Eighteenth Conference on Computational Natural Language Learning. CoNLL 2014, Ann Arbor, Michigan (2014), https://doi.org/10.3115/v1/W14-1609
26. Pennington, J., Socher, R., Manning, C.: Glove: Global Vectors for Word Representation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. EMNLP 2014, Doha, Qatar (2014), https://www.aclweb.org/anthology/D14-1162/
27. Plank, B.: Neural Cross-Lingual Transfer and Limited Annotated Data for Named Entity Recognition in Danish. In: Proceedings of the 22^{nd} Nordic Conference on Computational Linguistics. NoDaLiDa 2019, Turku, Finland (2019), https://www.aclweb.org/anthology/W19-6143
28. Santos, D., Freitas, C., Gonçalo Oliveira, H., Carvalho, P.: Second HAREM: New Challenges and Old Wisdom. In: Computational Processing of the Portuguese Language, 8^{th} International Conference, Proceedings. PROPOR 2008, Aveiro, Portugal (2008), https://doi.org/10.1007/978-3-540-85980-2_22
29. Santos, D., Seco, N., Cardoso, N., Vilela, R.: HAREM: An Advanced NER Evaluation Contest for Portuguese. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation. LREC 2006, Genoa, Italy (2006), http://www.lrec-conf.org/proceedings/lrec2006/pdf/59_pdf.pdf
30. Steingrímsson, S., Helgadóttir, S., Rögnvaldsson, E., Barkarson, S., Gudnason, J.: Risamálheild: A Very Large Icelandic Text Corpus. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation. LREC 2018, Miyazaki, Japan (2018), https://www.aclweb.org/anthology/L18-1690
31. Tjong Kim Sang, E.F., De Meulder, F.: Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In: Proceedings of the Conference on Computational Natural Language Learning. CoNLL 2003, Edmonton, Canada (2003), https://www.aclweb.org/anthology/W03-0419
32. Tryggvason, A.: Named Entity Recognition for Icelandic. Research report, Reykjavik University (2009)
33. Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., Pyysalo, S.: Multilingual is not enough: BERT for Finnish. arXiv e-prints arXiv:1912.07076 (2019)
34. Weegar, R., Pérez, A., Casillas, A., Oronoz, M.: Recent advances in Swedish and Spanish medical entity recognition in clinical texts using deep neural approaches. BMC Medical Informatics and Decision Making **19** (2019)
35. Wu, Y.C., Fan, T.K., Lee, Y.S., Yen, S.J.: Extracting Named Entities Using Support Vector Machines. In: Bremer, E.G., Hakenberg, J., Han, E.H.S., Berrar, D., Dubitzky, W. (eds.) Knowledge Discovery in Life Science Literature. pp. 91–103. Springer Berlin Heidelberg, Berlin, Heidelberg (2006), https://doi.org/10.1007/11683568_8
36. Yadav, V., Bethard, S.: A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. In: Proceedings of the 27^{th} International Conference on Computational Linguistics. COLING 2018, Santa Fe, NM, USA (2018), https://www.aclweb.org/anthology/C18-1182.pdf