# Named Entity Recognition with FLAIR Using Biomedical Corpora: Model Training and Evaluation

Bachelor Thesis Exposé

Volha Taliaronak
Supervisor:
March 2019

Humboldt Universität zu Berlin
Institut für Informatik

## Introduction

In the modern world, large volumes of biomedical texts are produced, collected, and stored, and as a consequence efficient ways to analyze and manage these data are required.

Recognition and classification of biological and medical terms or entities, such as genes, proteins, diseases, and other, in written language are very difficult due to many factors, including complexity of biomedical nomenclature, general lack of naming conventions, excessive use of abbreviations, frequent usage of synonyms and homonyms, and the fact that biological objects often have names consisting of many single words, digits, hyphens, other characters (Buzhou Tang et al., 2014; Leser et al., 2005). The problem of management and extraction of such information in a constructive way has obtained careful attention from researchers and companies (e.g., Google, Facebook, and other).

## Related Work

Over the last two decades, in order to improve the quality and efficiency of the information extraction process (IE), substantial research efforts have focused on the area of Natural Language Processing (NLP). IE deals with finding and extracting relevant information about defined entities and ignoring other information. One of the fundamental tasks of IE is Named entity recognition (NER). It is a process of identifying and classifying pieces of information that refer to a specified entity in the structured or unstructured texts.

Biomedical named entity recognition (BNER) is a core technique used to identify and categorize information about proteins, genes and other biomedical entities under pre-defined classes in a given text. Some of the early BNER methods relied on multiple manually defined rules. But, in the last decade, this approach has been abandoned because of its complexity and high time consumption. Today, due to the growth of computer power, a growing number of BNER systems are based on machine learning and hybrid methods.

Current NER methods rely on pre-defined features that try to capture the specific surface properties of entity types, properties of the typical local context, background knowledge, and linguistic information (Habibi et al., 2017). These features must have a machine-readable form – numeric or vector representation. Word Embedding, also called distributed representation, is a set of techniques, that represent words as vectors. Such vector representations are able to capture a large number of semantic properties and linguistic relationships between words.

Some researchers (Akbib et al., 2018) divide the current word embeddings methods into three embeddings types:

1. Classical word embeddings, pre-trained over very large corpora and shown to capture latent syntactic and semantic similarities.
2. Character-level features, which are not pre-trained, but trained on task data to capture task-specific subword features.
3. Contextualized word embeddings that capture word semantics in context to address the polysemous and context-dependent nature of words.

Word Embeddings is an active research area. There are many branches and research groups working on word embeddings, trying to figure out better data representations than the existing ones. Most new word embedding techniques rely on neural network architecture and unsupervised learning.

Word2Vec is a group of neural embedding models, which includes Continuous Bag-of-Words Model (CBOW) and Continuous Skip-Gram Model (CSGM). Word2Vec was introduced by (Mikolov te al., 2013) and has acquired popularity in 2013.

CBOW predicts a target word using the continuously distributed representation of its context and computing an average of this context. CSGM method relies on the idea that the closest words in context have more relation to a target word and must be weighted heavier than more distant words. The current word is used to predict the words within a pre-defined range.

Another commonly used model is Global Vector (GloVe) presented by (Pennington et al., 2014). It is a count-based model that collects co-occurrence statistics of words in the global corpus in the form of a co-occurrence matrix and weights all co-occurrences equally.

These models have already proved their effectiveness in practice, but there are still several directions for further improvements. Peters et al. (2018) introduce a new type of deep contextualized word representation that models both (1) complex characteristics of word use (e.g., syntax and semantics), and (2) how these uses vary across linguistic contexts (i.e., to model polysemy) and demonstrates that these representations can significantly improve the state of the art of some NLP problems (Peters et al., 2018).

In order to overcome the limitations of word representations, some research groups experiment with deep neural networks that learn character-based word representations (Cícero et al., 2014; Lample et al., 2016).

Moreover, one of the recent research lines focuses on the study of relation embeddings (Bouraoui et al., 2018).

As we can see, there still exist some relevant research areas explored yet.

## FLAIR Framework

The rapid increase in machine-readable data makes automatic information extraction much more attractive. There exist a lot of different Software for Text Analysis that interact with written language using deep learning algorithms and can give different outputs based on the learned required task.

FLAIR is a framework for state-of-art NLP, developed by Zalando Research[1].

This framework allows applying different NLP models, such as NER, part-of-speech tagging (PoS), sense disambiguation, and classification to the given text. It supports the usage of different embeddings, such as BERT embeddings, ELMo embeddings, Flair embeddings and combinations of them for training of new sequence labeling and text classification models. Flair embeddings method was developed and introduced by Zalando research group. This method is based on recent advances in neural language modeling and can generate a contextualized embedding for any string of characters in a sentential context (Akbik et al., 2018).

In this paper, we will focus only on the training of the Word Embedding Model for a biomedical NER task.

## Problem Statement

Evaluation of Flair framework on some well-known datasets, such as CoNLL-03, Ontonotes corpus, and other, shows high F1-score results in the sequence labeling tasks of NER for German and English languages. But these experiments do not involve biomedical corpora.

1 See: https://github.com/zalandoresearch/flair

**Goal of This Thesis**

The goal of this thesis is to train our own NER models for given biomedical corpora using Flair framework and evaluate the results.

**Approach**

Our work will consist of two steps:

1. We train our own corpus-specific NER models using Flair embeddings and parameter values suggested by the framework's developers. For models training, we use a collection of 25 BioNER corpora[1] that cover five biomedical domains and define five entity types, respectively:

1. Cell lines
2. Chemicals
3. Diseases
4. Genes
5. Species

Each corpus is split into training, development, and test sets in the ration 6:1:32. Each entity-domain contains five corpora. For each small NER model, we use one corpus with one pre-defined entity type.

2. We test trained NER models using test-data and report the results. To evaluate the performance of the framework we use standard metrics: Precision, Recall, F1-score.

1 See: https://github.com/leonweber/ner_scripts

## References


1. Alan Akbib, Duncan Blythe, Roland Vollgraf. 2018. "Contextual String Embeddings for Sequence Labeling". In Proceedings of the 27th International Conference on Computational Linguistics, (COLING 2018). Retrieved February 01, 2019, from http://aclweb.org/anthology/C18-1139.

2. Buzhou Tang, Hongxin Cao, Xiaolong Wang, Qingcai Chen, and Hua Xu. 2014. "Research Article Evaluating Word Representation Features in Biomedical Named Entity Recognition Tasks", Hindawi Publishing Corporation BioMed Research International, vol. 2014, Article ID 240403, 6 pages. Retrieved February 18, 2019, from http://dx.doi.org/10.1155/2014/240403.

3. Cícero Noguerira dos Santos, Bianca Zadrozny. 2014. "Learning Character-level Representations for Part-of-Speech Tagging". In Proceedings of the 31th International Conference on Machine Learning, PMLR 32(2); 1818-1826. Retrieved February 28, 2019, from http://proceedings.mlr.press/v32/santos14.pdf.

4. Gullaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, Chris Dyer. 2016. "Neural Architectures for Named Entity Recognition". Retrieved February 10, 2019, from https://arxiv.org/abs/1603.01360

5. Jeffrey Pennington, Richard Socher, Christopher D.Manning. 2014. "GloVe: Global Vectors for Word Representation". In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1532-1543. Retrieved February 10, 2019, from https://www.aclweb.org/anthology/D14-1162.

6. Maryam Habibi, Leon Weber, Mariana Neves, David Luis Wiegandt, and Ulf Leser. 2017. "Deep learning with word embeddings improves biomedical named entity recognition", Bioinformatics, 33, 2017, i37-i48, doi: 10.1093/bioinformatics/btx228.

7. Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, Luke Zettlemoyer. 2018. "Deep contextualized word representations". 6th International Conference on Learning Representations. Retrieved February 13, 2019, from http://www.aclweb.org/anthology/N18-1202.

8. Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean. 2013. "Efficient Estimation of Word Representations in Vector Space". arXiv:1301.3781v3 [cs.CL] 7 Sep 2013.

9. Ulf Leser, Jörg Hakenberg.2005. "What makes a gene name? Named entity recognition in the biomedical literature", Briefings in Bioinformatics, vol.6, no.4, pp. 357-369. Retrieved January 06, 2019, from https://academic.oup.com/bib/article/6/4/357/499223.

10. Zied Bouraoui, Shoaib Jameel, Steven Schockaert. 2018. "Relation Induction in Word Embeddings Revisited". In Proceedings of the 27th International Conference on Computational Linguistics, (COLING 2018). Retrieved February 14, 2019, from http://aclweb.org/anthology/C18-1138.