# Literature review for XCS224u: project gNER

**Matthias Droth**
matthias.droth@gmail.com

**Martin Nigsch**
martin@nigsch.eu

**Vasco Ribeiro**
vascosousaribeiro@yahoo.com

## Abstract

In our project gNER, we assess how to efficiently perform Named Entity Recognition (NER) in the German language. We assess models, in-domain and generic pretraining, word representation and tokenization choices and to a minor extent annotation strategies with the overall goal of finding an optimum between maximizing extraction accuracy while minimizing annotation and computational costs. Questions we will explore are: is domain specific or colloquial language pretraining more efficient, if any? How large are the incremental benefits of using state-of-the-art models, measured by number of training examples needed to pass a certain accuracy extraction threshold? What is the effect of word representations which are available to us? Are there clever concepts like adding more sentence or document contextual information to increase NER efficiency? The present literature overview provides the background to the project team in order to address those questions while building on state of the art knowledge.

## 1 General problem/task definition

We seek to extract named entities out of German text at a minimum overall cost. In order to do so, there are several areas to explore: (i) better models making better use of training data annotation, (ii) pretraining with colloquial vs. domain specific language, (iii) model choices beyond the obvious hyperparameter optimization (e.g. tokenization), and (iv) advantageous annotation strategies where the marginal benefit of adding more annotations is the greatest.

While we acknowledge the potential of cleverly chosen annotation strategies, we think that this area is the most time consuming and the hardest to implement within the time available for this project. For that reason, we focus on comparing NER efficiency resulting from the choice of algorithms as well as their pretraining.

## 2 Concise summaries of the articles

### 2.1 Trafilatura: A Web Scraping Library and Command-Line Tool for Text Discovery and Extraction Barbaresi (2021)

#### 2.1.1 Problem description

There is a range of problems that rely on text scraped from the web. In particular, procuring a domain-specific corpus might require scraping; even more, a large domain-agnostic general language corpus for pretraining requires a much larger corpus.

The presented tool (Trafilatura) covers text discovery and retrieval. In particular, it recognizes amongst a large mass of HTML codes which section of an unknown web page is the one containing the relevant textual information. The generic solution of this problem is neither obvious nor straightforward.

#### 2.1.2 Why this paper

Our project is based on NER on scraped news articles from a German speaking newspaper. This paper makes sense to explore to have an idea on the state of the art regarding web scraping and as well, to be informed on usual performance assessments and expected benchmarks.

#### 2.1.3 Methodology

A modular software package is presented which is based on a pipeline of filters and content heuristics. The key elements of this pipeline are

- Content delimitation based on HTML tags: selected nodes of the DOM tree are checked for relevance by element type, text length and link density and handed over to the next step.

- A fallback to catch faulty extractions is applied. This fallback algorithm is based on line length, text-to-markup ratio and position of nodes in the DOM tree.
- A further fallback is applied which looks for text elements that still might have been missed, e.g. based on div tokens.

With those elements, main texts are returned along with metadata. A language detection using CLD3 is applied optionally. The goal of these pipelines is to retain text and discard unwanted clutter.

## 2.2 A Survey on Recent Advances in Named Entity Recognition from Deep Learning Models Yadav and Bethard (2018)

### 2.2.1 Problem description

This paper presents a survey of research in Named Entity Recognition (NER) with a particular emphasis on comparing results achieved by more recent approaches.

### 2.2.2 Why this paper

With this survey, the authors aim to establish a taxonomy of efforts in this space in order to then provide pointers to most promising categories of modelling approaches.

### 2.2.3 Methodology

To determine the universe of relevant research the authors identified relevant keyword searches to perform on Google, Google Scholar and Semantic Scholar. Top 3 papers in each search category by citation count where read and considered for the survey whenever the paper introduced a new neural architecture or when it was a top performer on an NER dataset. The survey starts with knowledge-based and unsupervised bootstrapped systems moving on to a description of both classic feature-engineered supervised systems as well as modern feature-inferring neural network models. Drilling in on the latter the paper establishes the taxonomy as word-level, character-level and hybrid (character+word and, finally, char+word+affix). The vast majority of architectures presented rely on bi-directional LSTM layers coupled with Conditional Random Fields for final classification. The paper then proceeds to summarize the results obtained by each type of model on standard NER datasets enabling an overall ranking to be determined (in order of increasing performance) as follows: 1. Feature-engineered systems (least performant even despite using domain-specific rules, knowledge, features and lexicons); 2. Word-based and char-based models; 3. Char+word based hybrid models; 4. Char+word+affix hybrid models. The char+word+affix model showcases the potential of combining useful past insights with modern neural network models which, the paper concludes, represents an exciting avenue for future research.

## 2.3 Neural Architectures for Named Entity Recognition Lample et al. (2016)

### 2.3.1 Problem description

In 2016, state-of-the-art NER systems still were relying heavily on hand-crafted features. The use of neural architectures was promising to rely less on those features and, like so, have more generic models that lead to less costly deployment of NER adoptions to new domains.

### 2.3.2 Why this paper

This article is an often-cited base article introducing BiLSTMs in conjunction with conditional random fields. It explains the base concepts and provides a wide range of performance benchmarks, including in particular German examples which can provide benchmarks for our own project. Similarly, it provides quantitative guidance on the effects of pretraining and also of adding character-level word embeddings which is the real claimed step forward that the authors claim to have achieved with this paper.

### 2.3.3 Methodology

Two neural architectures are presented: (i) a model called LSTM-CRF consisting of a bidirectional LSTM and a conditional random field which has the desired property of modelling tagging decisions jointly. (ii) a transition-based chunking model for NER.

In (i), a bidirectional LSTM is set up with word representations obtained by concatenating their left and right contexts. Subsequently, a conditional random field is used to model tagging decisions jointly, i.e. by explicitly considering that those depend on each other. Adding character-based features proves to be positively impacting NER performance. Dropout during training is providing a further boost.

In (ii), a previously published model called Stack-LSTM is re-used. The LSTM-CRF model seems to be more robust and requires less ortographic information. The authors hypothetize that this is because the bidirectional LSTMs provide more contextual information.

## 2.4 The annotated transformer Rush (2018)

### 2.4.1 Problem description

Reimplementing models from published papers is common practice within the open-source NLP community but often also a struggle. As a consequence, (i) it makes it difficult to improve scores on established benchmarks and (ii) it poses a pedagogical issue as students face a barrier reproducing results from scientific literature. The challenges of reproducing published results are often amplified by the need for extensive hyperparameter tuning and long training times.

### 2.4.2 Why this paper

The authors of this literature review face the challenges just described when considering to employ a transformer architecture in the course project. Rush's *The annotated transformer* article interweaves text from the original transformer publication [Vaswani et al. (2017)] with easy to understand code, thus giving the authors confidence to move forward with transformers for the course project.

### 2.4.3 Methodology

The overall goal of the article is to facilitate understanding of the transformer architecture by showing (i) text from the original publication and (ii) images of according code implementations next to each other. Most text passages are verbatim repetitions of the original passages. Differences exist, however, as some sections have been reordered or deleted.

A fully working implementation is available under `https://nlp.seas.harvard.edu/2018/04/03/attention.html`.

## 2.5 Bidirectional LSTM-CRF Models for Sequence Tagging Huang et al. (2015)

### 2.5.1 Problem description

Sequence tagging (which includes among others, part-of-speech tagging (POS), chunking and named-entity recognition (NER)) is an important NLU problem. The paper further argues for sequence tagging as being useful in the context of other pipelines, such as determining that a user's search engine query includes a product as part of its text.

### 2.5.2 Why this paper

This paper presents a detailed introduction to popular neural architectures for sequence tagging problems. Among them are (uni and bi-directional) LSTM (Long-Short Term Memory) with the additional possibility of a CRF (Conditional Random Field) layer for the named-entity classifications.

### 2.5.3 Methodology

The authors train word-level LSTM, BI-LSTM, CRF, LSTM-CRF and BI-LSTM-CRF models and test them on three benchmark NLP tagging tasks (POS, chunking and NER). The BI-LSTM-CRF model outperforms the others as well as a third-party Convolutional Neural Network (CNN) - CRF model. Additionally the authors find less dependency of the BI-LSTM-CRF model on word embedding initialization when compared with the CNN-CRF model.

## 2.6 Simple and Effective Few-Shot Named Entity Recognition with Structured Nearest Neighbor Learning Yang and Katiyar (2020)

### 2.6.1 Problem description

Few-shot named entity recognition remains a challenging problem. While algorithms exist, the method presented in the paper improves performance relative to state-of-the art algorithms. The problem is relevant because training data for NER is very expensive to obtain.

### 2.6.2 Why this paper

This paper contributes to understand annotation-efficient NER. We think that the present method might not be directly applicable as it is based on English NER pretainings. However, the approach presented might have advantages.

In addition to that, the present paper provides as well link to further state-of-the-art models and hence also gives interesting links for further study.

### 2.6.3 Methodology

The base idea of the paper is to train a NER model first on four publicly available datasets. Then, label dependencies are captured using a nearest neighbor classifier and a Viterbi decoder. Effectively, the most similar entity out of the set the algorithm was initially trained with is picked.

This algorithm is then compared with a BiLSTM NER and a BERT-based NER. Although the authors claim that the results are state-of-the-art, the average F1 scores are still low.

## 2.7 Hierarchical Contextualized Representation for Named Entity Recognition Luo et al. (2019)

### 2.7.1 Problem description

The BiLSTM algorithm for named entity recognition algorithm is in widespread use, however does not incorporate global sentence-level or document-level information which has the potential to increase NER accuracy and efficiency. This paper explores the effectiveness of an algorithm which adds this contextual information in a hierarchical way.

### 2.7.2 Why this paper

This paper fits in nicely with the overall target to explore how to increase NER efficiency in a new domain, new language setting: exploiting information that is globally available like document and sentence embeddings seem promising avenues for our problem, too with the potential to ultimately get higher accuracy NER results for the same effort put into annotation.

### 2.7.3 Methodology

A word-level and character-level embedding based on IntNet is the base word representation. This representation is the fed into a BiLSTM. A label representation in the same embedding space as the word embeddings is derived. This is done by using an attention mechanism, optimizing for closeness between NER label representations and the representations of the target words. Subsequently, CNNs are used to capture relative spatial information among consecutive words. A sentence level representation is obtained via averaging the hidden states of the BiLSTM mentioned above. This sentence representation is appended to the base word representation as mentioned above. Further, a document-level representation is derived which a key-value memory network whose design is inspired from an original proposal from the domain of Question Answering: the keys correspond to questions, and the values to answers. Again, with an attention algorithm, a document level representation is derived and, after combination with the word and sentence level representation, fed into a conditional random field decoder. The authors claim to exceed state-of-the-art NER performance (article released in November 2019).

## 2.8 German's Next Language Model Branden Chan (2020)

### 2.8.1 Problem description

While most of the available pretrained models are available in English, the present paper explores the computationally efficient creation of the official BERT and ELECTRA models for German language and makes the result available for general purpose re-use.

### 2.8.2 Why this paper

In order to solve the problem of annotation-efficient NER in German language, we need an overview on state-of-the-art NER models in German language. While the present paper is focused on training German language models in a computationally efficient way, the results also provide a good overview on state-of-the-art models and how they were trained.

Further, as the models are uploaded to the hugging face model hub, we might re-use the models derived in this paper in the actual project.

### 2.8.3 Methodology

The work describes the use of 4 different German language datasets for pretaining in different sizes and mentions the inherent bias and quality issues due to their source. The models are used for both text classification and NER. The data for the latter task is GermEval14, one of the largest German NER datasets. The methodlogy used is optimized for early stopping and best use of scarce computational resources. An interesting conclustion was that the marginal use of extra data seems not to drastically increase model performance.

## 2.9 Dice Loss for Data-imbalanced NLP Tasks Li et al. (2020)

### 2.9.1 Problem description

In the presence of NLP tasks with significant data imbalances (such as Named Entity Recognition (NER) in which most tokens fall under a catch-all class) traditional accuracy-focused loss metrics (such as Cross-Entropy (CE) loss) are overly affected by easy negatives. This leads models to struggle learning from the crucial hard negatives as well as the positive examples.

### 2.9.2 Why this paper

Using three different loss functions that attempt to more equally account for false positives and false negatives (and are hence more immune to data imbalance issues) the authors compare performance on standard NLP problems with standard CE-loss based benchmarks.

### 2.9.3 Methodology

The authors introduce Dice Loss (DL), Focal Loss (FL) and Dice Coefficient (DSC) loss metrics and compare BERT-based models using these metrics with traditional CE loss. A variety of different problems (such as PoS tagging and NER) and baselines are contrasted with models using the 3 loss metrics. Consistent outperformance of models using these 3 loss metrics is observed when compared to models using the CE loss metric. Further tests with positive results are carried out by training models on datasets constructed in such a way as to have different degrees of imbalances.

## 3 Compare and contrast

We selected the articles for this literature review such that we would have a solid basis for the work laid out in the task definition.

Our problem can be briefly described with the following sequence.

1. Scrape articles
2. Label named entities – with possible efficiency gains via active learning and use of gazetteers if possible
3. Apply NER models, varying pretraining, representation of words choice, and model itself
4. Assess results

The goal of the literature review was thus to get to a comprehensive overview on scraping, state-of-the-art NER models and techniques, a deep-dive on fundamental models as well as individual promising routes like the choice of the loss function or also the value of pretraining for a specific domain or language. A possible different interesting additional focus would have been the effectiveness of different annotation strategies, in particular active learning schemes. For the purposes of this project, though, we believe that a focus on pure models is broad enough, hence the exclusion of annotation strategy exploration. Further (even though we didn't find many) we didn't consider domain-specific papers to be hugely important as our NER use-cases rely on quite generic categories (such as organizations, addresses, etc.).

In this spirit, Barbaresi (2021) forms a natural start as it explores web scraping. For our problem, this has several applications: first, as a means to procure the base data. Second, in case we would want to construct a larger language corpus for pretraining, the tool mentioned in the article could be a good base choice.

Then, as a conceptual base, it made sense for us to include a review of articles which maps out a lay of the land of current approaches in NER: Yadav and Bethard (2018). Besides knowledge of standard model approaches in this space we chiefly retain from this article NER evaluation metrics. It would be interesting to understand to which extent the metrics used in other articles might differ in their precise definitions from the ones outlined in this article. However, this isn't covered in any of our reviewed articles in sufficient precision, so we cannot make a statement here. From the approaches presented, we deduce from the presentation in the article that we will focus our further interest on supervised systems learning to replace human created curated rules. Neural networks generally outperform feature-engineered supervised systems. Hence, we zoom in purely on feature-inferring neural network systems which are anyway outlined in the article as the clear state-of-the-art these days. Further, we take the idea of gazetteers out of this article to be used for speeding up the labeling process.

The core of the literature review is then formed by a range of articles describing individual NER algorithms: Huang et al. (2015) introduces two core algorithms in the sequence-tagging space: BiLSTMs and CRFs Lample et al. (2016) extends these by considering both word-level and character-level inputs simultaneously, relying also on BiLSTMs and CRFs as well as a transition-based approach. In Rush (2018), the transformer architecture is really well explained; for us, this formed the educational basis to dive as well into transformer-based models which seem to be promising performance-wise. With Luo et al. (2019), a further step up in terms of model complexity and as well performance was achieved: the approach to add contextualized representations to the ”simple” BiLSTM model is promising and credible to improve NER model performance. The approach taken by Yang and Katiyar (2020) is interesting: we understand this as optimizing for label embeddings to be close to word embeddings; then unknown entities are extracted by extracting the most similar known label that the model has been trained for. Understandable, this approach can't perform as well, but has the advantage of needing really few training examples (few shot learning).

The comparison is rounded up with a view on German pretraining Branden Chan (2020): GBERT and GELECTRA are two German language models released by Chan, Schweter and Möller at deepset.ai. This paper is interesting also because it focuses on training practicalities that allow to optimize the output given limited computational budget.

Finally, the exploration of a further modelling choice opens up an interesting avenue: we rarely question the omnipresent use of the cross-entropy loss objective. In Li et al. (2020), the described ”dice loss” claims to achieve significant performance boost without changing model architectures.

## 4 Future work

These papers do not contradict each other; much rather, in our view their combination and assessment on one common dataset already presents an interesting project. The axis we would like to explore is to assess the mentioned algorithms on one common, non-English (German) dataset.

The practical aspect we would like to explore is to make experiments in the sense that we artificially limit the number of training examples that the models see (e.g. 5/20/50/75/100) and to check then the performance as precision and F1 score by algorithm.

Then, a grid can be filled by varying the following dimensions:

- Effects of domain-agnostic generic pre-training
- Effects of domain specific language pre-training
- Effects of more labelled training data

For a practitioner interested in tangible guidance whether to rather invest in model choice, pretraining on a domain specific corpus, or a larger initial labelling corpus, we are convinced that such a comparison would be highly appreciated. In the assumption that with limited resources, not all axes can be optimized at once, such a guidance is practically very valuable and currently unknown to the authors of this paper.

## References

Adrien Barbaresi. 2021. Trafilatura: A web scraping library and command-line tool for text discovery and extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 122–131, Online. Association for Computational Linguistics.

Timo Möller Branden Chan, Stefan Schweter. 2020. German's next language model. *Accepted by COLING2020*, arXiv:2010.10906. Version 4.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991.

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*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. 2020. Dice loss for data-imbalanced NLP tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 465–476, Online. Association for Computational Linguistics.

Ying Luo, Fengshun Xiao, and Hai Zhao. 2019. Hierarchical contextualized representation for named entity recognition. *CoRR*, abs/1911.02257.

Alexander Rush. 2018. The annotated transformer. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 52–60, Melbourne, Australia. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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*, pages 2145–2158, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yi Yang and Arzoo Katiyar. 2020. Simple and effective few-shot named entity recognition with structured nearest neighbor learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6365–6375, Online. Association for Computational Linguistics.