Data Driven Enrichment of Historical Low-Resource Languages for Foundational NLP Tasks and their Neural Network Models

Dissertation

zur Erlangung des Doktorgrades
der Naturwissenschaften

vorgelegt beim Fachbreich Informatik und Mathematik
der Johann Wolfgang Goethe-Universität
in Frankfurt am Main

von
Sajawel Ahmed
aus Frankfurt am Main

Frankfurt, 2023
(D 30)

Vom Fachbereich (12) Informatik und Mathematik der
Johann Wolfgang Goethe-Universität als Dissertation angenommen.

Dekan: Prof. Dr. Martin Möller

Gutachter: Prof. Dr. Gemma Roig und Prof. Dr. Alexander Mehler

Datum der Disputation: 15.06.2023

Sajawel Ahmed
Data Driven Enrichment of Historical Low-Resource Languages for Foundational NLP Tasks and their Neural Network Models
Dissertation, 2023

Goethe-Universität Frankfurt
Institut für Informatik
Robert-Mayer-Str. 11-15
60325 Frankfurt, Deutschland

Acknowledgement

This work would have not been possible without the continuous help of countless people involved in it and around me during the long-enduring time of my PhD studies. First, and foremost, I would like to thank Prof. G. Roig for her support and supervision during my PhD studies, providing me regularly new Machine Learning perspectives, guiding me with writing well-developed research papers, advising me with designing the main structure of this cumulative dissertation, and motivating me throughout the time in a pleasant way. I highly appreciate her trust in my abilities and the independence she gave me, leading to interesting contacts and prolific collaborations with various research institutes around the world. On the same page, I would like to thank Prof. A. Mehler for his support and supervision during my PhD studies, under whom I received a foundational training of the current Natural Language Processing research and its publication process, and started to understand the world of academia and its surroundings. Especially his structured and elegant way of approaching complex scientific problems provided me the necessary guidance starting from the very first day as a researcher.

I would like to pose my sincere gratitude to Prof. Ö. Özsoy, Dr. M. Rehman and further colleagues from the Faculty for Linguistics, Cultures, and Arts for their trust in my expertise and the valuable research work we delivered together. Furthermore, I am incredibly grateful to Prof. R. V. Zicari for being my mentor throughout my PhD studies, providing me overall guidance even in times of uncertainties and obscurities. Similarly, I am also grateful to Prof. L. Hedrich and additional colleagues from the institute for their support and friendly talks, especially during the final period of this dissertation. A special thanks go to Prof. R. v. d. Goot, Prof. B. Plank and further colleagues from the ITU Copenhagen for the great research collaboration and the productive time I could have while being a visiting researcher at their lab.

Last but not least, I would like to thank all my friends who supported me throughout the time, and most importantly my beloved family and siblings, S. Ahmed, H. Ahmed A. Ahmed, M. Ahmed and R. Ahmed, and especially my parents, without whom this work would have never been possible—thank you, this work is dedicated to you, my dear family!

Abstract

In the recent past, we are making huge progress in the field of *Artificial Intelligence*. Since the rise of neural networks, astonishing new frontiers are continuously being discovered. The development is so fast that overall no major technical limits are in sight. Hence, digitization has expanded from the base of academia and industry to such an extent that it is prevalent in the politics, mass media and even popular arts. The DFG-funded project *Specialized Information Service for Biodiversity Research* and the BMBF-funded project *Linked Open Tafsir* can be placed exactly in that overall development. Both projects aim to build an intelligent, up-to-date, modern research infrastructure on *biodiversity* and *theological studies* for scholars researching in these respective fields of historical science. Starting from digitized *German* and *Arabic* historical literature containing so far unavailable valuable knowledge on biodiversity and theological studies, at its core, our dissertation targets to incorporate state-of-the-art *Machine Learning* methods for analyzing natural language texts of *low-resource languages* and enabling foundational *Natural Language Processing* tasks on them, such as *Sentence Boundary Detection*, *Named Entity Recognition*, and *Topic Modeling*. This ultimately leads to paving the way for new scientific discoveries in the historical disciplines of natural science and humanities. By enriching the landscape of historical low-resource languages with valuable annotation data, our work becomes part of the greater movement of digitizing the society, thus allowing people to focus on things which really matter in science and industry.

Contents

1 Introduction

“The scholars and nations of the past which have ceased to exist, were constantly employed in writing books about various fields of science and wisdom, regarding those that were to come after them, and anticipating for a reward proportionate to their ability, and trusting that their endeavors would meet with acknowledgment, attention, and remembrance—content as they were even with a small degree of praise; small if compared with pains which they had undergone, and the difficulties which they had encountered, in revealing the secrets of science and its obscurities.”

— Algorismi, *Liber Algebrae et Almucabola*

Language is part of human diversity—modern technology should preserve and promote this diversity, rather than reducing it. In our modern times, we are making huge progress in the field of *Artificial Intelligence* (AI). Since the rise of neural networks, astonishing new frontiers are continuously being discovered. The development is so fast that overall no major technical limits are in sight. In *Natural Language Processing* (NLP), the majority of research work is conducted in English, a high-resource language, for which a large amount of previous work and resources are available. This definitely accelerates the progress of the ongoing *big data* driven NLP, such as it is visible in various NLP benchmarks (e.g. SNLI [Bow+15] for *Natural Language Inference*, IIRC [Fer+20] for *Machine Reading Comprehension*, SQUAD [RJL18] for *Question Answering*). From the perspective of research on AI, it is indeed beneficial to continue the research on a language, which already has reached a high level of digitization. This will bring us even more quickly toward reaching the goal of developing some form of human-like *strong AI*. However, from a societal and ethical point of view, this mere focus on one language for the sake of other *existing* languages is not justified, given the rising need for NLP models of non-English backgrounds. As a consequence, a gap is created between modern, high-resource and historical, low-resource languages.

In the past ages, there were many historical languages that were important for various parts of human society and their activities. These used to be lingua franca of science, arts, commerce, and everyday life. Languages, such as Ancient Egyptian, Ancient Greek, Classical Arabic, or Premodern German (with its Fraktur script), which possess large volumes of historical literature, were and still are to this date

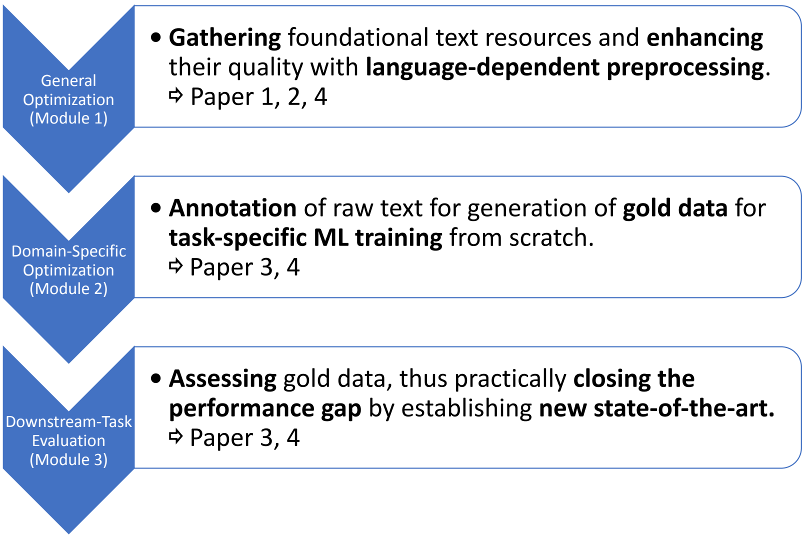

General Optimization (Module 1)
• **Gathering** foundational text resources and **enhancing** their quality with **language-dependent preprocessing**.
➪ Paper 1, 2, 4
Domain-Specific Optimization (Module 2)
• **Annotation** of raw text for generation of **gold data** for **task-specific ML training** from scratch.
➪ Paper 3, 4
Downstream-Task Evaluation (Module 3)
• **Assessing** gold data, thus practically **closing the performance gap** by establishing **new state-of-the-art.**
➪ Paper 3, 4

Fig. 1.1.: Overview of contributions along thematic modules and their corresponding study papers.

relevant for many (scientific) communities and (religious) societies, lay their foundations and even shape their future development. With the demise of these old societies and their replacement with modern civilizations, only some parts of their cultural heritage were carried forward. The majority was thus buried in handwritten manuscripts and printed books, of which some have survived until our current age. These important historical languages with their large treasures deserve the attention of current and ever increasing NLP research. In order to perform historical analysis which are relevant for our modern age, we need to allow these *forgotten* low-resource languages to benefit from the wave of machine learning (ML) progress, thus making historical texts accessible to modern scientific studies and ethically approaching an egalitarian state of NLP research.

In our dissertation, we close the growing resource and performance gap by analyzing step by step two particularly different low-resource languages, namely Premodern German for the domain of biodiversity literature, and Classical Arabic for the domain of theological literature. We do this according to the examples of foundational NLP tasks such as *Sentence Boundary Detection* (SBD) [SA19], *Named Entity Recognition* (NER) [ASM18; Ahm+19; Ahm+22], and *Topic Modeling* (TM) [Ahm+22]. By doing this, we deploy the following modular procedure, which specifies our major contributions as well. Figure 1.1 gives an overview of these.

Module 1 We perform a *general optimization* in respect to models developed for high-resource languages. In this optimization, we gather all available open-source resources for our target languages and enhance their basic textual quality with sophisticated preprocessing methods, which are specifically developed after analyzing the language and its script in the historical context [ASM18; SA19; Ahm+22]. In this way, we create a bridge between the state-of-the-art models (which are predominantly designed for English) and the target language.

Module 2 We perform a *domain-specific optimization* in respect to models developed for high-resource languages. In this optimization, we create labeled training datasets by manually annotating the historical books of our target languages (German: *BIOfid* corpus on the biodiversity of plant, birds, moths and butterflies [Ahm+19], Arabic: *Tafsir Al-Tabari* books on exegetical studies of law, ethics and philosophy [Ahm+22]). By choosing NER as our target task, we generate gold data with over 15k and 51k sentences for German and Arabic, respectively, the first domain-specific datasets yet for these languages. This in turn allows us to train robust task-specific language models from scratch and utilize the full potential of the existing state-of-the-art models.

Module 3 We combine both Module 1 and Module 2 by performing a first *downstream-task evaluation* of our newly generated dataset with the novel language-dependent preprocessing methods. By reaching a final performance on our chosen NLP task close to that of the state-of-the-art performance for English on the same task, we thus practically close the gap between these high- and low-resource languages [Ahm+19; Ahm+22].

Although our target languages and their domains are quite different in nature, we see in our work that, from an ML perspective, these points are not relevant; rather, their language-specific grammatical structure and writing systems are. Thus, our work facilitates an automatic extraction of historical information that has been buried so far in the bulk of paper manuscripts and volumes. By creating the necessary training data for tackling foundational NLP tasks (such as SBD, NER or TM) with various current ML algorithms, we provide an open-source gold standard for the NLP community and hereby lay the foundations for future work on the digitization of historical studies.

1.1 Background and Motivation: Low-Resource Languages

In NLP, the term *low-resource language* denotes those human languages which do not have a sufficient amount of digitized text resources. These can either be *unlabeled data* (i.e. raw text, e.g. digitized books, papers, online media), which can be used to train general language models (LM), such as *Word2vec* [Mik+13] or *BERT* [Dev+19], or *labeled data* (i.e. structured text, e.g. dictionaries, treebanks, databases), which are specifically annotated by domain-experts to be used for a task-specific training of neural models in combination with pre-trained LMs, such as the *BiLSTM* model [Lam+16] for NER with Word2vec embeddings.

Low-resource languages stand in contrast to high-resource languages, like English or Chinese, which can be used to work with the most recent ML models directly, as they have sufficient amount of resources open-source available. However, as the needs of the data-intensive models change constantly, there is as yet no standard threshold which would determine to which class a given language belongs. Be that as it may, most historical languages, including our target languages, are low-resourced and require a sophisticated treatment by the current NLP research.

1.1.1 Stages of Human Language Development

Since the invention of writing, human languages have gone through many important transitions in the courses of their individual histories. In the beginning, stones, metal, clay tablets, and even bones were used to write down (mostly short) pieces of human language texts. In the next stage, these heavyweight materials were replaced by papyrus, parchment, and ultimately paper, whose lightweight nature allowed for a rapid spread of written communication, thereby immensely increasing literacy among the general human population. In our modern times, we are in the midst of another important transition, in which new digital technologies are continuously revolutionizing means and ways of communication [Li+20]. More and more pieces of text are written out in the "weightless" digital format, thus making the requisite of a physical medium secondary or even obsolete.

Every transitional step has eased the access to written human language and its semantic content. However, in every step, language diversity was lost as well. Those societies and their languages that could not cope with the changing trends and developments began to decline and ultimately became extinct (see Figure 1.2). The existence of low-resource languages and their future treatment should be viewed from this perspective. Having learned from this historical development, the aim of researchers should be to aim for the least possible loss in this modern step of *digital transition*, and preserving the language diversity and its culture as best they can, thus introducing an egalitarian state of AI research [Zic+21].

As human individuals, it is important to know ourselves and the history of our ancestors in order to make informed decisions on where to go individually in the future. Similarly, as a human civilization, it is important for us to understand

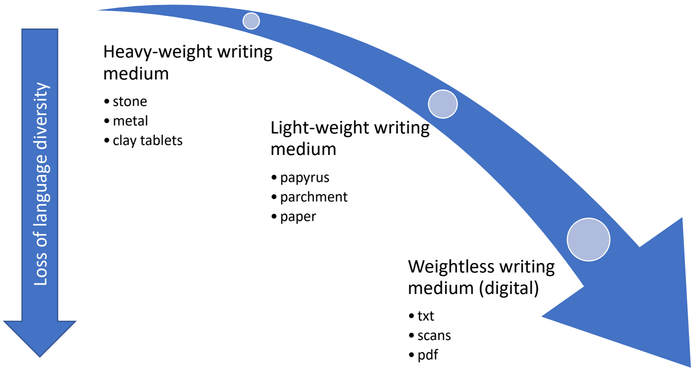

Loss of language diversity
Heavy-weight writing medium
• stone
• metal
• clay tablets
Light-weight writing medium
• papyrus
• parchment
• paper
Weightless writing medium (digital)
• txt
• scans
• pdf

Fig. 1.2.: Major stages of human languages developments since the advent of writing. Along every transitional step, the access to written language becomes easier, however, the overall diversity of languages decreases.

past societies and their evolution to the stage where we are today, to understand specifically how they were thinking, working and living in former times, in order to decide where to go together in the future. The current advances in NLP allow us to capture the semantics of societies going back to the time of the advent of first written documents, a time which is so far filled with myths and facts. Thus, our foundational research work allows to better understand portions of human history and enables a revival of forgotten knowledge, even from "dead" languages.

1.2 Research Questions

In this cumulative dissertation, we investigate the area of historical NLP primarily through the example of the foundational NLP task of NER. Starting our research work with German NER, and closing it with the Arabic counterpart, the following four major research questions (RQs) are addressed by our study papers:

- **RQ-1:** How is it possible to close the NER performance gap between high- and low-resource languages in the the domain of historical literature?

- **RQ-2:** Is an annotation data-driven optimization without any network designing sufficient for low-resource languages, or is it necessary to create adjusted architectural designs for each of them?

- **RQ-3:** What is the role of the text data size of language resources for downstream-task evaluations (e.g. NER)?

- **RQ-4:** What is the role of the script size of language resources and their syntactical structure for downstream-task evaluations (e.g. NER), in case features, such as diacritic points and vocal markers, are removed from their specific scripts (e.g. for Classical Arabic)?

By answering these questions paper by paper, we demonstrate for our chosen languages that the cost-intensive annotation process is necessary for the digitization of historical literature and its further analysis by modern NLP methods. We show that the generation of *annotation data* is essential for overcoming the state of being a low-resource language, and provide overall guidelines to NLP researchers undertaking the same endeavors.

1.3 Structure of the Dissertation

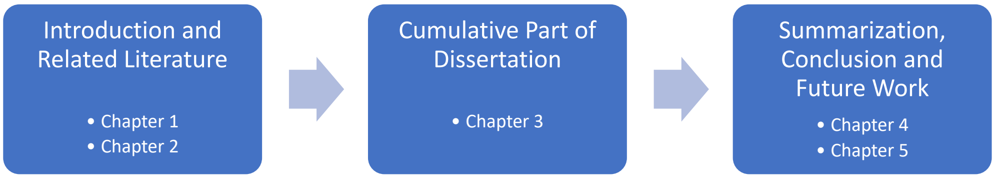

Introduction and Related Literature
• Chapter 1
• Chapter 2
Cumulative Part of Dissertation
• Chapter 3
Summarization, Conclusion and Future Work
• Chapter 4
• Chapter 5

Fig. 1.3.: Overview of the main structure of this dissertation.

Our cumulative dissertation consists of three major parts, as shown in Figure 1.3. The structure of the remaining dissertation is as follows:

Chapter 2 provides an overview of related literature. Firstly, it gives details on the technical background of the ecosystem of the NLP models, in which we navigated throughout our study papers. Secondly, in light of these technical developments of recent years, it presents the previous work on low-resource languages, especially in respect to open-source available data resources.

Chapter 3 consists of the main cumulative part of this dissertation in which each sub chapter corresponds to the respective study paper (cf. Figure 1.1). The first three papers belong to the analysis on German NER, where in *Paper 1* [ASM18] a general optimization of NER is performed, in *Paper 2* [SA19] a general optimization of SBD is conducted, and in *Paper 3* [Ahm+19] both are utilized to perform the final domain-specific optimization for historical biodiversity literature. *Paper 4* [Ahm+22] represents the Arabic counterpart, which performs all of these steps at once and thus generalizes this procedure.

Chapter 4 summarizes the contributions of our dissertation, and creates an overall procedure for transforming a given low-resource language to a higher state.

Chapter 5 finally concludes the dissertation by discussing the contributions, the possibilities and limits of automatizing the transformation procedure, and giving directions to future research work, especially on the future treatment of low-resource languages both in academia and industry.

2 Related Literature

2.1 Ecosystem of NLP Models

"(Our work has been possible only by) standing on the shoulders of previous giants."

— Newton

Since the rise of neural networks, there has been a growing number of NLP models appearing freely online on platforms such as *GitHub*[1] or *HuggingFace*[2] which have been pre-trained on large text collections. Within our data driven framework, we have made extensive usage of such existing pre-trained models. Due to their large model parameter size, a training from scratch was not always possible for us given the limited access to compute resources we had, especially in respect to the most recent models based on *transformers* [Vas+17], for instance *BERT* [Dev+19], *XLNet* [Yan+19] or *GPT-3* [Bro+20]. This limitation led us to focus on a data driven enrichment of historical low-resource languages for foundational NLP tasks, such as NER. In other words, we focused only on the data input side of the (predominantly English) models, without performing any architectural designing of the neural networks.

We present the major components of neural network models from our dissertation that have allowed us to achieve noteworthy progress on NLP for historical low-resource languages. At the beginning, these neural models consisted of two separately trained components: 1) foundational LMs (*static word embeddings*), modeling the general knowledge from large unlabeled text corpora; and 2) task-specific neural networks, modeling the domain knowledge from labeled training data, such as for NER. Starting from 2019, these components were joined into one large model (namely BERT) which contains both 1) foundational LMs (*dynamic word embeddings*), and 2) a last neural network layer which can be fine-tuned task-specifically, such as for NER. In the next sections, these components are presented briefly.

[1] `https://github.com/`
[2] `https://huggingface.co/`

2.1.1 Word2vec (Static Word Embeddings)

Starting with the pioneering work of Bengio et al. [Ben+03], the concept of word embeddings and its further development for NLP tasks saw a continuous rise within the research community. Until recently, the model of *Word2vec* [Mik+13] and its further extensions [PSM14; LG14; KM16; Boj+17] were the foundations of most ongoing research in NLP with neural networks. Based on the context of a given piece of text, the model embeds words, phrases or sentences into high dimensional vector spaces. In such vector spaces, the associations of words, phrases or sentences are captured semantically in such a way that algebraic operations lead to reasonable relationships (e.g. on word level: $\mathrm{vec}(king) - \mathrm{vec}(man) + \mathrm{vec}(woman) \approx \mathrm{vec}(queen)$ [Mik+13]). This property has been immensely useful throughout our study papers. In addition to that, the Word2vec algorithm has a compute-friendly runtime, which allows researcher to train their individually adjusted models from scratch. This feature of the algorithm has been an important aspect within our low-resource scenario, especially for the Arabic counterpart of our dissertation, where we utilized a *full experimental setup*, in which both the unlabeled data for the training of the word embeddings and the labeled data for the training of the task-specific model are adapted according to the specific preprocessing method of *script-compression* [Ahm+22].

2.1.2 Long Short-Term Memory (LSTM) Network

For conducting specific supervised NLP tasks, such as NER, the general word embeddings require further processing in combination with task-specific datasets. In recent years, this has been done primarily with neural models consisting of *Long Short-Term Memory* (LSTM) networks [HS97] for text data. Central to their success has been their ability to memorize long distance relationships in time, hence allowing for a possible link between the first starting word of a book and its last concluding one. Yadav and Berthard [YB18] provide an overview of this model type on the example of NER, where a neural architecture of bi-directional LSTMs is utilized (BiLSTM). The *Word Representations* are pre-trained on large text corpora with the Word2vec algorithm, whereas the *Char Embeddings* are initiated randomly and optimized during the task-specific training of the BiLSTM model. Due to the relatively lightweight characteristics of this model, a training from scratch is possible, which has been an important aspect for our study papers along the compute-friendly runtime of the Word2vec algorithm.

2.1.3 Transformers and BERT (Dynamic Word Embeddings)

However, NLP researchers started to realize that there are constraints with the static word embeddings. Due to their static nature, they are not able to consider the context of words and phrases in context-sensitive environments [CLS14; Tia+14; McC+17]. Human languages are characterized by exactly such dynamic environments, and among other complexities, they contain a varying degree of words with *polysemy* (i.e. one word has many meanings). With the work on *ELMo* [Pet+18] and especially *BERT* [Dev+19], the obstacle of the polysemy problem was resolved. Now, for each word of an input sentence, a *contextualized word embedding* is given by the pre-trained LM, which can be different for the same word in another sentence. The BiLSTM-based *ELMo* and later the transformer-based *OpenAI GPT* started with an unidirectional objective of language modeling; *BERT* fully developed this potential by considering both left and right contexts with a neural architecture of deep bidirectional transformers [Dev+19; Wan+20].

This extensive extension ultimately joined both separated components described in previous sections into one large model, which led to a major progress in the field of NLP, especially for cross-lingual scenarios. In such scenarios, the *compute-intensive* transformer-based models also allowed multilingual *transfer learning* [DR19] (i.e. the case where there are few labeled examples for our target language available, but the model has been intensively trained on the same task with multiple other languages and consequently can transfer its learning to the target language). Thus, in a very short period of time, BERT became "*a ubiquitous baseline in NLP experiments*" [RKR20] after its open-source publication, leading to various new research studies re-analyzing and improving existing NLP tasks and their datasets. However, the heavyweight nature of these BERT models hampers for most NLP researchers the possibility of training individually adjusted BERT models from scratch. In consequence, their research endeavors are limited by the availability of pre-trained models, unless they have access to strong compute infrastructures such as those found at larger research institutes and big tech companies [LD21].

Aspects of both the lightweight and heavyweight models have been key to the research work in our dissertation. Hence, depending on the specific situation, we considered the lightweight scenario in our NLP ecosystem by training individually adjusted word embeddings from scratch in combination with LSTM-based models, once no pre-trained BERT models were available.

2.2 Previous Work on Low-Resource Languages

Until 2017, most practitioners in NLP were working mainly for the English language. Even for modern languages, such a Standard High German, not many resources were available [ASM18]. Only after the advent of large-scale pre-trained language models like *ELMo* [Pet+18], and especially *BERT* [Dev+19], *XLNet* [Yan+19], *GPT-3* [Bro+20], much progress had been made for modern low-resource languages. Many researchers started to create various variants of multilingual models (e.g. *mBERT* [Dev+19], *XML-R* [Con+20]) and uploaded their pre-trained versions open-source on *Hugging Face* [Wol+20], allowing for cross-lingual transfer learning and zero-shot learning on languages with limited resources.

However, seen from the perspective of historical NLP, the benefits of these developments are limited. With the support of large online-available unlabeled text corpora (e.g. Wikipedia[3]), most of these multilingual LMs were built for modern languages and their genre of online media. Firstly, this makes them suitable only for unsupervised learning tasks. For supervised learning tasks (e.g. NER), annotation data is still required to actually perform a fine-tuning of the LM and to create an adapted version of it [Ahm+19; Ahm+22]. Secondly, this makes them suitable only for modern texts. Despite the arising possibility of transfer learning along the temporal variants of a given language (e.g. mBERT trained on Modern Standard Arabic, applied on Classical Arabic), our study papers have shown that this situation is not ideal, as linguistically these are two different languages with two different genres. Hence, these pre-trained LMs are only applicable if there are no better ones available for the historical language and its specific genre. As a consequence, the requirements for more work on resource generation for historical NLP still remains. There is a significant need for more research work on low-resource languages. There are for instance researchers studying dead languages. Given the current situation, they cannot leverage the advantages of the ongoing NLP research. Thus, they are bounded by traditional (mostly manual) research methods of the humanities.

In the next section, we will present an overview of historical low-resource languages along their existing resources, thereby shedding light on the question of which other historical languages have been digitized, before proceeding to our two languages of focus in the domains of biodiversity and theology.

[3] https://www.wikipedia.org/

2.2.1 Historical Low-Resource Languages

There are numerous historical low-resource languages, which possess large treasures of written past knowledge and therefore deserve the attention of current NLP research. For comparative reasons, we have prepared a list of major historical languages (independent of domain) which already have some primary digitized resources available, and for which further annotation work can be conducted in order to make them useful for general historical studies.

Language	Language Model	Task & Dataset
Ancient Greek	*Ancient-Greek-BERT* [SRL21]	POS [SRL21]
Biblical Hebrew	*BEREL* (BERT) [Shm+22]	HD [Shm+22]
Classical Chinese	*AnchiBERT* [Tia+21]	NER [Wan+21]
Egyptian (Coptic)	Word2Vec [ZMT20]	DP, NER [ZMT20]
Latin	*LatinBERT* [BB20]	NER [Erd+19]
Sanskrit	*Sanskrit-BERT* [San+21]	SD [San+21]
Sumerian	*Sumerian-BERT* [Ban+21]	NER [WLH22]

Tab. 2.1.: Overview of historical low-resource languages and their open-source available pre-trained language models along task-specific training datasets (POS: Part-of-Speech tagging, HD: Homograph Disambiguation, DP: Dependency Parsing, SD: Synonym Detection).

As the list in Table 2.1 is quite small-scaled and limited mainly to low-level NLP tasks, we can see that the development is still in its infancy; however, some of the mentioned historical languages actually have a potential to overcome the digital transformation quickly, as more digital resources are regularly being published open-source, waiting to be used and further analyzed by NLP researchers. Especially for Latin, an ancient language which has been important for the history and culture of most Western European nations, we have various further resources available, as there are research institutes which have been working on this language even prior to the rise of neural networks in NLP (e.g. *Latin Word2vec* (static) and *Latin Flair* (dynamic) [Sto+20], *Frankfurt Latin Lexicon* [Meh+20], *TTLab Latin Tagger* for POS tagging [Meh+15], *Patrologia Latina* corpus in the *eHumanities Desktop* [Meh+10; Gle+09]).

2.2.2 German Biodiversity Literature

German was one of the major languages of science and arts up until the modern era and has retained this position in many scientific communities even until today. In the past, various researchers wrote in German on the topic of biodiversity of the Central European regions. Much printed literature had been produced by these old researchers. However, as of now, there are not many domain-specific resources available for German biodiversity NER or any related primary NLP tasks, which would enable an automatic analysis of these historical literature. Our updated review has shown that apart from the creation of the *BIOfid Dataset* (cf. *Paper 3* [Ahm+19]), and its extensive follow-up work as part of the *Specialized Information Service for Biodiversity Research*[4] [Lüc+21], there are still no related biological NER datasets available. A researcher aiming to develop an NER model for an application in this specific field of historical studies has currently only one option, unlike its English counterpart, which provides more options with the *LINNAEUS* [GNB10], *Species-800* [Paf+13], *COPIUS* [NGA19], or *BiodivNERE* [Abd+22] datasets.

2.2.3 Arabic Theology Literature

In contrast to the above mentioned research line, there is a growing number of works related to the Arabic language. The reason is primarily the rising interest of researchers and scholars from the vast area of Arabic speaking countries (with over 620 million L1 and L2 speakers, respectively). Still, these resources are primarily built for Modern Standard Arabic for the domain of news and online media, e.g. the Arabic part of the *mBERT* model with the support of Arabic Wikipedia, or *ANER* [BRB07] and *AQMAR* [Moh+12] datasets for NER with the support of newspaper articles. On the other hand, there are some NER datasets available for Classical Arabic, such as *Bedaya Corpus* (1,161 sentences) [MS21], *CANERCorp* (N/A) [SZ18], *NoorCorp* (3,818 sentences) [SM17]. However, these datasets are relatively small compared to the ones for other high-resource languages (e.g. English *ConLL-2003 NER* dataset with over 22,137 sentences [TD03]), lack the target CoNLL format with sentence boundaries, and are in their entirety not open-source available. Hence, apart from our large-scale annotation work on *Tafsir Dataset*[5] (cf. *Paper 4* [Ahm+22]), this historical language in general and its domain of theology specifically still require more research work on resource generation, which would be available open-source and thus trigger further progress in the field of historical theological studies.

2.3 Perspectives

There has been a growing number of research works which have started to consider *modern low-resource languages* and make their training datasets and pre-trained models freely available. This allows various researchers to use them directly for downstream-task evaluations and practical applications. We can see that there are some parallel developments in this progress. On the one side, we observe a strong boost of technological advancement with the introduction of large transformer-based

[4] www.biofid.de

[5] https://www.tafsirtabari.com/

models. Due to their large model parameter size, they are able to successfully capture more text data during their training. This in turn allows researchers to extend the models by new training paradigms (e.g. transfer learning), new text genres (e.g. science, law, history, art, social media), and especially new low-resource languages (e.g. multi-lingual training). On the other side, we observe a rise of awareness among NLP researchers for this sensitive topic of low-resource languages [Rud+21]. More people have access to neural networks and their open-source software frameworks (e.g. *Tensorflow*[6], *Keras*[7], *PyTorch*[8]), so more researchers have started to understand the *digital needs* of speakers of other languages apart from English, and have started working towards enriching the landscape of low-resource languages. Now, the question, which development came first (technological advancement or rise of awareness), might be similarly answered, as the question, if the egg came first or the chicken.

Nevertheless, regarding the *historical low-resource languages*, the development is still in its infancy. There is a lack of funding for research into these languages, and consequently, there is a lack of modern ML models and datasets available open-source, which can be directly applied by historical researchers from the field of humanities. In the next chapter, the main part of this dissertation, we will see a demonstration of how step by step, these goals can be achieved for such low-resource languages.

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

[7] `https://keras.io/`

[8] `https://pytorch.org/`

3 Cumulative Part of Dissertation

Resource-Size matters: Improving Neural Named Entity Recognition with Optimized Large Corpora

Sajawel Ahmed, Manuel Stoeckel, Alexander Mehler
Text Technology Lab
Goethe University Frankfurt
Frankfurt, Germany
{sahmed,mehler}@em.uni-frankfurt.de

***Abstract*—This study improves the performance of neural named entity recognition by a margin of up to 11% in F-score on the example of a low-resource language like German, thereby outperforming existing baselines and establishing a new state-of-the-art on each single open-source dataset. Rather than designing deeper and wider hybrid neural architectures, we gather all available resources and perform a detailed optimization and grammar-dependent morphological processing consisting of lemmatization and part-of-speech tagging prior to exposing the raw data to any training process. We test our approach in a threefold monolingual experimental setup of a) single, b) joint, and c) optimized training and shed light on the dependency of downstream-tasks on the size of corpora used to compute word embeddings.**

***Index Terms*—named entity recognition, word embeddings, lemmatization, part-of-speech, neural networks, nlp**

I. Introduction

Named Entity Recognition (NER) is a crucial part of various *Natural Language Processing* (NLP) tasks like entity linking, relation extraction, machine reading and ultimately *Question Answering* (QA). With the recent rise of neural networks, much emphasis has been put on high-resource languages like English or Chinese leading to fast advancements of many foundational tasks, in particular NER which in many areas reaches near-human performance for these languages [1], [2]. However, for other, less-resource languages like German, their neural NER counterparts did not attract similar attention from the deep learning community, leading to lower performance by a margin of up to 11% F-score.

In this paper, we look for the reasons and take steps towards solving them. By example of German we bridge the current gap between the performance of neural NER for different languages and bring the performance to a new state-of-the-art. We report evidence that the inferior quality of German text data and its small size are the major reasons for the observed lack of progress.

To tackle this problem, we use a larger corpus for training the foundational word embeddings, namely *Leipzig40* [3] (including the whole German Wikipedia until 2016) combined with the *WMT 2010 German monolingual training data* [4], and contrast its use with the *COW corpus* [5], the largest collection of German texts extracted from web documents with over 617 Mio. sentences. Besides, we bring all scattered (open-source) resources of annotated NER datasets for German together which are to date available, prepare and merge them to increase the amount of the final training data. This includes the major NER datasets of *CoNLL-2003* [6] and *GermEval-2014* [7], and the smaller datasets of *Europarl-2010* [8] and of *EuropeanaNewspapers-2016* [9]. To this collection, we add the dataset of Tübingen Treebank (*TüBa-D/Z*) [10], which to the knowledge of the authors is utilized the first time for the task of neural NER.

It is an increasing scientific practice to make models open source accessible. New models appear almost daily, for example in the *Deep Learning* (DL) community. As a consequence, changing existing models and trying out different hybrid setups is getting a scientific practice involving more and more scientists. This is advantageous, since attempts to improve existing models can contribute to their validation. However, it is often forgotten that *data is the gold of scientist*: it is the availability of limited resources that leads to significant improvements in various areas such as CoNLL, SNLI [11] and SQuAD [12] for the tasks NER, *natural language inference* and QA and stand behind the recent success of neural networks in NLP. Therefore it is important to consider sufficient available resources, to annotate them according to the task and to optimize them if necessary. This task is often time-consuming and costly. The present paper deals with assessing the impact of resources to NER by example of a rather low-resource language like German. We show the influence of different training sets on the performance of neural NER, of different combinations of these data sets and above all of different levels of their preprocessing. We deal with the aspect of resource optimization with regard to lemmatization and *Part-of-Speech* (POS) tagging and analyze their influence besides the training of word embeddings and task-specific neural networks. Our main finding is: an increase of size and quality of the (task-independent) word embedding corpus and of the (task-specific) training dataset leads to a significant improvement of sequence labeling tasks like NER, which can be larger than just an amendment of the underlying neural architecture. For the future of neural NER by example of less- or low-resource languages this means: collecting unlabeled corpora for training morphology-dependent, high quality embeddings is a good alternative to increase the performance of downstream-tasks.

The remainder of the paper is organized as follows: Section 2 reviews related work, Section 3 presents a sketch of the underlying model, Section 4 describes our threefold experimental setup of a) single, b) joint, and c) resource optimized training, Section 5 reports and discusses our results, and, finally, Section 6 draws a conclusion.

II. Related Work

Compared to high-resource languages, comparatively less emphasis has been put on the task of neural NER by example of German. Noteworthy work has been done so far only by [13] on GermEval and by [1] on CoNLL; both will be used as baselines here. Reimers et al. [13] were among the first to apply neural networks to German NER. However, they did not consider GermEval in combination with CoNLL. Apart from them, the remaining studies (predominantly conducted by non-native speakers) consider this task as a side product of dealing with various other languages. In this way, the state-of-the-art on German neural NER has been established by [1] in 2016.

Gillick et al. [14] consider German as a variant in a multilingual training setup while additionally considering the datasets of two Germanic languages (English and Dutch) and one Romanic language (Spanish) from the CoNLL shared task; as a result, they reach 76.22 % F-score. However, for the single training on the German part of CoNLL they stay below [13].

From the point of view of resource optimization, the recent work of [15] is worth mentioning. Klimek et al. also observe the gap between the languages and therefore carry out a detailed analysis of the difficulties for the German NER task using the GermEval data set as an example. They come to the conclusion that *"the task of German NER could benefit from integrating morphological processing"* [15]. To this end, we start our analysis and apply our designed morphological processing approach to all text corpora and NER datasets.

III. Model

Our neural model consist of two separately trained components: a) foundational word embeddings, modeling the general knowledge from large unlabeled text corpora, and b) task-specific neural networks, modeling the domain knowledge from the labeled training data. In this section, both components are presented briefly.

a) Word Embeddings: The language model of continuous space word representations (*word2vec*) [16] and its variations by [17], [18] are the foundations of most ongoing research in NLP with neural networks. Based on the context, the model embeds words, phrases or sentences into high dimensional vector spaces. In such a space, the semantics of associations of words and phrases are captured to such an extent that algebraic operations lead to meaningful relationships (e.g. vec(*king*) − vec(*man*) + vec(*woman*) ≈ vec(*queen*) [16]). This property is immensely useful for our application. We use the model of *word2vec* and its extension *wang2vec* [19] which explores syntactic data and, thus, better suites the task of NER.

b) Neural Model: We give a brief sketch of the neural model *LSTM-CRF* which we use throughout this paper. The model is similar to the one used in [1], which goes back to the works of [20]–[22]. We use a neural model consisting of stacked LSTM and CRF layers. The *base layer* is made of two parts: (i) a preprocessing sublayer generating the character-based embeddings with a cell of forward and backward LSTMs (*biLSTM*) [23], and the word embeddings from the input sentence, (ii) followed by an encoding sublayer again with a cell of a biLSTM extracting features and generating compressed hidden representations. The *prediction layer* is made of CRFs and takes the previous hidden representations to finally produce the *Named Entity* (NE) tag predictions.

Let $(w_1, \dots, w_{N_s}) = [w_i]$ be the list of words of a sentence from the input corpus of texts. Furthermore, let $(c_{i,1}, \dots, c_{i,N_{w_i}}) = [c_{i,l}]$ be the list of characters of the word w_i consisting of N_{w_i} characters with $c_{i,l}$ being its l^{th} character. For a given word w_i and its NE-tag (gold label) $t_i \in \{$*PER, LOC, ORG, MISC, O*$\}$ the data flow within the neural network is as follow:

$$\text{char2vec}(c_{i,l}) \mapsto \vec{c_{i,l}} \tag{1}$$

$$\text{biLSTM}([\vec{c_{i,l}}]) \mapsto \vec{h_i^c} \tag{2}$$

$$\text{word2vec}(w_i) \mapsto \vec{w_i} \tag{3}$$

$$\text{biLSTM}([(\vec{w_i}, \vec{h_i^c})]) \mapsto [\vec{h_i^w}] \tag{4}$$

$$\text{CRF}([\vec{h_i^w}]) \mapsto [t_i] \tag{5}$$

where char2vec is a (randomly initialized) lookup table for embedding all characters into a corresponding vector space, and $(\vec{w_i}, \vec{h_i^c})$ is the concatenation of the embedding vector of word w_i and its character-based hidden representation. The model is trained to predict the NE-tag t_i for each word after seeing the whole input sentence at once.

IV. Experimental Setup

A. Datasets

In order to evaluate our model of Section III for neural NER on German data, we put emphasis on the major datasets of CoNLL (German part) and GermEval. However, more German resources are available that have so far gone unnoticed in the DL community. In Table I, we gather all these NER datasets, which are to date freely accessible, and list them along their number of sentences. Additionally, for each dataset the total number of NE tokens is provided along the four categories from the standards defined in the CoNLL shared task 2003 (CoNLL format). Table I shows that the TüBa-D/Z dataset is the largest of these, both in terms of the number of sentences and of tokens, ideally fitting to the needs of deep neural networks.

TABLE I
NER Datasets

Corpus	***Sent.***	***PER***	***LOC***	***ORG***	***MISC***
CoNLL-2003	18,024	**8,309**	7,864	7,621	4,748
Europarl-2010	4,395	514	724	874	**966**
GermEval-2014	31,300	16,204	**16,675**	12,885	9,254
Europ.Newsp.-2016	8,879	**7,914**	6,143	2,784	3
TüBa-D/Z-2018	**104,787**	**55,746**	28,582	32,224	12,865

a) Preprocessing of Training Data: Apart from CoNLL, most copora had to be further processed to fit the CoNLL format. For GermEval, we consider only the top-level NE, refraining from nested NE to stay in line with the remaining datasets. As a tagging scheme, we preferred the BIO (IOB2) scheme, as it has been shown to perform better [24]. All datasets are given in the BIO scheme, except CoNLL (IOB1) and Europarl (IOB1), which we converted into the target scheme.

For EuropeanaNewspapers, we take the two datasets written in standard German orthography, namely *enp_DE.lft.bio* and *enp_DE.sbb.bio* based on historic newspapers from the Dr. Friedrich Tessmann Library and the Berlin State Library, respectively, and omit the Austrian historic newspapers which use a different orthography, differing heavily from the former samples. The original dataset is not provided in the 4-column CoNLL format, which writes each word of a sentence horizontally along its lemma, POS tag and NE-label, and separates each sentence by an empty newline. Therefore, we convert the data into our target format by using *spaCy V2.0*[1] which by its recent release supports preprocessing German texts by providing language models for sentence boundary detection, lemmatization and POS tagging.

For TüBa-D/Z, we extracted the NE-tags from the *tuebadz-11.0-conll2010* version. In the case of nested NE, we use a filtering heuristics to extract the longest spanning NE, which allowed us to get more robust training data, not splitting well known entities into parts (e.g. *[Goethe Universität Frankfurt]*_ORG vs. *[Goethe]*_PER *Universität [Frankfurt]*_LOC). We converted the tagging scheme of TüBa-D/Z to our target format. Lastly, to allow comparisons with other NER datasets, we mapped the NE category *Geo Political Entity* (GPE) to *LOC*.

b) Data Splitting & Merging: For CoNLL and GermEval we use the splits as provided in the original datasets. Further, we split TüBa-D/Z into train/dev/test sets according to the common ratio of 80/10/10 percentages. Due to the smaller size of the Europarl und the EuropeanaNewspapers datasets, we did not consider them for the first experimental setup of single training, rather we merged them with the training data for the second experimental setup of joint training. For this setup, we aligned all datasets by mapping the NE category *OTH* to *MISC* to fit to the CoNLL format. In this way, we generated the currently largest training dataset for German NER of a size of $133,258$ sentences.[2]

B. Word Embeddings

German is a highly inflected language compared to English or Chinese whose syntax is more analytic. For languages like German, the embedding of a single word (e.g. *klein*) is dispersed across its various morphological and spelling variants (stem: *klein* → *kleiner*, *kleinste*, *kleine*, *kleines*, *kleinen*, *kleinem*, *Klein* etc.), therefore reducing the number of its samples and weakening its information value if not being lemmatized appropriately. On the other hand, languages with a rather analytical syntax show such morphological variants to a lesser extent, if at all. We assume that this difference is the reason why their embeddings are of higher quality and therefore their performance in downstream tasks is many times higher than in less analytical languages. In order to mitigate this factor for the German language in its negative effect, we are therefore forced to use embeddings of higher quality. In the experimental setup of single training, we tackle this by using more text data. Table II lists the corpora we use for training our word embeddings. Leipzig40-2018 contains the largest possible extract from the so-called Leipzig Corpora Collection in 2018, which was generated by its maintainers on demand for our study, omitting any possible duplicate sentences. To increase the corpus size we combine this extract with WMT-2010-German forming our so-called *LeipzigMT* corpus. Besides, we consider the COW-2016 corpus, arguably the largest text collection for German. This corpus contains not only a textbook-like language, as found for example in Wikipedia. Therefore, we assume that it fits well with the NER datasets used here, which in turn come from various sources (news, web, wikis, etc.). Both corpora are already preprocessed and split into sentences, containing words, numbers and punctuations. We do not remove punctuation marks, but separate them from words and numbers by surrounding them with spaces to avoid the introduction of variations with punctuation marks. In addition, as a preprocessing step, we write all words in lowercase to account for spelling and morphological variations.

TABLE II
TEXT CORPORA

Corpus	*Sentences*
Leipzig40-2018	40.00 Mio.
WMT-2010-German	19.36 Mio.
COW-2016	617.28 Mio.

In a third variant of our experiment we deepen the optimization of resources by taking into account lemmatization and POS tagging in connection with writing words in lower case. While lemmatization increases the observation frequency of words, POS tagging allows a more correct specification of their syntactic roles in sentences and consequently differentiates individual observations that are included in the calculation of embeddings. On the other hand, lower case writing of words removes ambiguities, as they are induced in German especially by capitalization at the beginning of sentences. Table III shows the variations we use for this setup.

We apply lemmatization and POS tagging in combination with writing words in lowercase to all resources before they are used in training. These conversions are coupled with an exact conversion of the NER data sets in the respective experiment to avoid mismatches and to increase the overlap with the trained embeddings. Again, we use spaCy for these tasks and use its

[1] http://spacy.io

[2] CoNLL (12,152) + GermEval (24,000) + Europarl (4,395) + EuropeanaNewspaper (8,879) + TüBa-D/Z (83,832)

language models for lemmatization and POS tagging. Listing 1 shows an example of this approach.

Listing 1. Example for Lemma & POS

```
raw sentence   : Kleine Kinder sind mutiger.
lemma          : Klein Kind sein mutig .
lemmapos       : Klein_ADJA Kind_NN sein_VAFIN mutig_ADJD ._$
lemmapos_lower: klein_ADJA kind_NN sein_VVFIN mutig_ADJD ._$
```

These conversions are intended to standardize any text input and thus to solve the above-mentioned problems in connection with morphological variations.

TABLE III
EMBEDDING VARIANTS PER EXPERIMENTAL SETUPS

Experimental Setup	***Variant***	***Features***
Single Training	1	lower
Joint Training	1	lower
Optimized Training	2	lemma
	3	lemma_lower
	4	lemmapos
	5	lemmapos_lower

C. Training Parameters

To remain comparable with the baseline models on CoNLL [1] and GermEval [13], we train the word embeddings with dimension 100[3], window size of 8 and minimum word count threshold of 4, consequently, setting the LSTM dimension to 100 as well[4]. We choose dimension 25 for character-based embeddings and the final CRF-layer, and train the network in 100 epochs with a batch-size of 1 and dropout rate of 0.5. As an optimization method, we use the stochastic gradient descent with a learning rate of 0.005. Apart from fitting the LSTM dimension to 300 while using the 300-dimensional pretrained German fastText embeddings [25], the model is fixed throughout our experiments to these settings. Any further sophisticated hyperparameter tuning (e.g. *Population Based Training*) is left for future work.

V. RESULTS

In this section, we present the results we obtained for our three experimental settings. As described in [24], we perform every experiment up to 6 times, starting from different random seeds, in order to arrive at significant final values on the respective test dataset. We evaluate the NER results by using the official evaluation script from the shared task of CoNLL 2003. All our experiments were run on Nvidia's *GTX 1080 Ti* GPUs.

A. Single Training

We compare our results with the current top performing models on CoNLL and GermEval. Table IV shows the highest results we achieve on the single training setup (first experimental setting).

TABLE IV
SINGLE TRAINING

Data	**Embeddings**	**Features**	***F-score*** **[%]**
CoNLL	*pre-trained* ***Leipzig***	wang2v	78.76 [1]
GermEval	*pre-trained* ***UKP2014***	word2v	75.9 [13]
CoNLL	*self-trained* ***LeipzigMT***	wang2v	80.81
CoNLL	*self-trained* ***COW***	wang2v	**83.29**
GermEval	*self-trained* ***LeipzigMT***	wang2v	81.97
GermEval	*self-trained* ***COW***	wang2v	**83.14**
TüBa-D/Z	*self-trained* ***LeipzigMT***	wang2v	88.95
TüBa-D/Z	*self-trained* ***COW***	wang2v	**89.26**

We achieve an improvement throughout the datasets, outperforming all previous results on German neural NER, and establishing a new state-of-the-art on each of them. Increasing the corpus size by means of the LeipzigMT corpus displays a side-by-side performance increase on the CoNLL baseline. Increasing the corpus size further through the COW corpus gives us finally the best results on CoNLL. From this perspective, looking at the three data points for CoNLL (or GermEval), we observe a logarithmic growth of F-score as a function of the size of the underlying embedding corpus. Even larger corpora than the COW corpus are needed to further support this observation.

On the side of training data, we observe a similar but more powerful behavior. On LeipzigMT, the increase of training data size from CoNLL to GermEval, and then to TüBa-D/Z leads to an improvement of +1.16% and +6.98% in F-score. For COW this behavior re-emerges for TüBa-D/Z, closing the gap to high-resource languages like English, and almost crossing the 90% barrier on TüBa-D/Z. Besides, we see that the larger train dataset TüBa-D/Z does not heavily depend on the corpus size implying that it is beneficial to invest in annotation efforts.

We also find that wang2vec generally performs better than word2vec. This shows that a task-specific embedding algorithm is important (in our case taking into account the syntax for NER).

Last but not least, our experiments show that keeping information about capitalization can even downgrade the quality of word embeddings. Likewise, we observe that integrating capitalization information as an additional input feature to our neural network does not lead to better results. We assume that this is due to the inflectional morphology of German, according to which nouns are capitalized at the beginning, in contrast to English, where mainly proper names (named entities) are written in this way.

B. Joint Training

As a first step towards joint training, we report the best results for fastText embeddings and compare them to UKP2014 embeddings, only using the two datasets from the baseline models. Next, we approach the full joint setup and perform the training on all German NER datasets. Starting from the results of the last section, we consider only COW for this setup. Table V shows the top results for this setup.

[3] Lample et al. [1] use dimension 100 for English, but 64 for German. We increase this dimension to close the gap.

[4] For word2vec, we performed an extensive search on numerous embeddings with dimension values $(50, 100, 150, 200, 300)$ along with minimum word count threshold and window size values in the range of $[4, 200]$ and $[5, 10]$, respectively. However, no major differences were observed in the final results.

For fastText, we get the best results among all settings we examined (the results on single training were worse than for this setup). However, they are still below the ones with UKP2014, which themselves were trained with the original word2vec model back in 2014. This shows, that the fastText algorithm, being a promising extension of word2vec, does not suit well to our NER task, even though using a more informative vector space with 300 dimensions. Hence, we discard it for further experiments.

For COW, the transfer learning on a single task works well and the performance for CoNLL and GermEval are improved further, lying slightly above the single training values. It can be noted that the final performance is more directed towards the low performing values. We assume that it depends more on the datasets with the lower single training performance (who make with $\sim 37\%$ a large part of the joint training dataset), as due to the data merging additional variety is introduced to the final training dataset. This makes the tasks more difficult and brings it closer to a real-world scenario. Still, the slightly improved performance indicates that the neural network is generalizing, and successfully performing *task-related transfer learning on datasets*, i.e. the model is improving the same task on a heterogeneous dataset, given that it performs well on a single large homogeneous dataset.

Overall, the results are promising; they indicate that we have a good candidate for applying a jointly trained tagger to large resources where the availability of labeled data is scarce.

TABLE V
JOINT TRAINING

Data	**Embeddings**	**Features**	***F-score*** **[%]**
CoNLL+GermEval	*pre-trained* ***UKP2014***	word2v	78.06
CoNLL+GermEval	*pre-trained* ***fastText***	300dim	77.00
all	*self-trained* ***COW***	wang2v	**83.47**

C. Resource Optimization via Lemmatization & POS tagging

In this final setup of resource optimization, we examine various constellations. Table VI reports the corresponding list of results.

Intuitively, using POS tagged sentences for training word embeddings may appear to be unusual, however, the results show a different picture. We get results very close to the top performances of the previous sections. A common pattern across all experiments can be detected. The variation of lemmatization on COW constantly delivers top scores for the three major datasets, and even produces the highest value for CoNLL across all setups. Lemmatization performs comparatively better than lemmatization combined with POS tagging. This shows that dispersing the semantics of a given word across various roles it can take does not improve the quality of the final embeddings. Rather it is better to decrease the (redundant) varieties in the vector space by assembling in advance all morphological variants to a common base form, which only then is mapped to a common semantic vector. After lemmatization is performed, we can see that lower casing does not lead to a notable improvement. We assume that lemmatization already performs a good filtering of the raw text, making lower casing almost ineffective.

TABLE VI
OPTIMIZED TRAINING VIA LEMMA & POS

Data	**Embeddings**	**Features**	***F-score*** **[%]**
CoNLL	***LeipzigMT***	lemma	82.57
	LeipzigMT	lemma_lower	82.94
	LeipzigMT	lemmapos	81.22
	LeipzigMT	lemmapos_lower	81.20
	COW	lemma	**83.64**
	COW	lemma_lower	83.14
	COW	lemmapos	82.38
	COW	lemmapos_lower!	82.47
GermEval	***LeipzigMT***	lemma	82.53
	LeipzigMT	lemma_lower	82.47
	LeipzigMT	lemmapos	81.46
	LeipzigMT	lemmapos_lower	81.05
	COW	lemma	**82.87**
	COW	lemma_lower	82.53
	COW	lemmapos	81.96
	COW	lemmapos_lower!	81.38
TüBa-D/Z	***LeipzigMT***	lemma	88.50
	LeipzigMT	lemma_lower	88.27
	LeipzigMT	lemmapos	87.85
	LeipzigMT	lemmapos_lower!	87.83
	COW	lemma	89.08
	COW	lemma_lower	**89.24**
	COW	lemmapos	88.43
	COW	lemmapos_lower	88.02

Regarding the size of the corpus used for generating the word embeddings, we come to the conclusion, that lemmatization and POS tagging reduce the performance differences from previous sections which depended so far on the latter size. This confirms our assumption that the word2vec algorithm in its original form does not suit well to morphological rich languages. The results of this setup show that the values for LeipzigMT and COW now lie closer to each other, making the performance to some extent independent from the size of the embedding corpus. This is an important finding, giving rise to promising opportunities and applications for low-resource languages.

VI. CONCLUSION & FUTURE WORK

In this paper, we performed a far reaching study on neural NER by example of a low-resource language like German. The study focused on a monolingual experimental setup. Nevertheless, the improved results pave the way for related languages with similar characteristics as German.

There are various ways to improve existing neural models. Instead of just designing deeper and wider hybrid models, we showed the high importance of gathering and merging resources and how their careful optimization can eliminate the lack of progress. In particular, we found out that increasing the size and improving the quality of raw corpora for word embeddings by applying morphological processing like lemmatization & POS tagging leads to meaningful improvements. In addition, we demonstrated the effect of transfer learning by merging data sets for a joint training setup, which also produced good results and makes this approach a promising candidate for NER applications in the area of scarce resources of annotated data sets.

Overall, we conducted the first comprehensive research for the German NER on all existing training data sets and resources, including the study of common pre-trained embeddings such as fastText. In this context, we established a new state-of-the-art using all open source data sets for the German NER, which exceeds the 80% F-score limit for the German NER and closes the gap to other high-resource languages such as English.

For future work we plan to further refine the training process of word embedding and in particular to investigate how the performance of downstream tasks can become more independent of the size of embedding corpora using linguistic methods such as lemmatization and POS tagging. To this end, we intent to examine the recently published ELMo embeddings [26] for German. Finally, we will examine the role of the multilingual COW corpus for word embedding by example of other languages such as Dutch, French, Spanish and English.

Acknowledgment

This work was funded by the German Research Foundations (DFG) as part of the *BIOfid* project[5] (DFG-326061700). The source code and the trained embeddings are uploaded on GitHub[6] for the research community. Special thanks goes to G. Lample for his directions on the procedure for training the embeddings, to Prof. G. Heyer and F. Helfer for providing the extract of Leipzig40-2018 corpus, and last but not least to Dr. A. J. Davis for his valuable comments on the final manuscript.

References

[1] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "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*. Association for Computational Linguistics, 2016, pp. 260–270.

[2] L. Ouyang, Y. Tian, H. Tang, and B. Zhang, "Chinese Named Entity Recognition Based on B-LSTM Neural Network with Additional Features," in *International Conference on Security, Privacy and Anonymity in Computation, Communication and Storage*. Springer, 2017, pp. 269–279.

[3] D. Goldhahn, T. Eckart, and U. Quasthoff, "Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages," in *LREC*, 2012.

[4] C. Callison-Burch, P. Koehn, C. Monz, K. Peterson, M. Przybocki, and O. F. Zaidan, "Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation," in *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*. Association for Computational Linguistics, 2010, pp. 17–53.

[5] R. Schäfer, "Processing and querying large web corpora with the COW14 architecture," in *Proceedings of Challenges in the Management of Large Corpora 3 (CMLC-3)*, P. Baski, H. Biber, E. Breiteneder, M. Kupietz, H. Lngen, and A. Witt, Eds., UCREL. Lancaster: IDS, 2015.

[6] E. F. Tjong Kim Sang and F. De Meulder, "Introduction to the conll-2003 shared task: Language-independent named entity recognition," in *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*. Association for Computational Linguistics, 2003, pp. 142–147.

[7] D. Benikova, C. Biemann, and M. Reznicek, "NoSta-D Named Entity Annotation for German: Guidelines and Dataset," in *LREC*, 2014.

[8] M. Faruqui and S. Padó, "Training and Evaluating a German Named Entity Recognizer with Semantic Generalization," in *Proceedings of KONVENS 2010*, Saarbrücken, Germany, 2010.

[9] C. Neudecker, "An Open Corpus for Named Entity Recognition in Historic Newspapers," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, N. C. C. Chair), K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds. Paris, France: European Language Resources Association (ELRA), may 2016.

[10] H. Telljohann, E. W. Hinrichs, S. Kübler, H. Zinsmeister, and K. Beck, "Stylebook for the Tübingen treebank of written German (TüBa-D/Z)," 2012.

[11] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2015.

[12] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ Questions for Machine Comprehension of Text," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 2383–2392.

[13] N. Reimers, J. Eckle-Kohler, C. Schnober, J. Kim, and I. Gurevych, "GermEval-2014: Nested Named Entity Recognition with Neural Networks," in *Workshop Proceedings of the 12th Edition of the KONVENS Conference*, G. Faaß and J. Ruppenhofer, Eds. Universitätsverlag Hildesheim, October 2014, pp. 117–120.

[14] D. Gillick, C. Brunk, O. Vinyals, and A. Subramanya, "Multilingual language processing from bytes," in *HLT-NAACL*, 2016.

[15] B. Klimek, M. Ackermann, A. Kirschenbaum, and S. Hellmann, "Investigating the Morphological Complexity of German Named Entities: The Case of the GermEval NER Challenge," in *Language Technologies for the Challenges of the Digital Age*, G. Rehm and T. Declerck, Eds. Cham: Springer International Publishing, 2018, pp. 130–145.

[16] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[17] O. Levy and Y. Goldberg, "Dependency-Based Word Embeddings." in *ACL (2)*, 2014, pp. 302–308.

[18] A. Komninos and S. Manandhar, "Dependency Based Embeddings for Sentence Classification Tasks." in *HLT-NAACL*, 2016, pp. 1490–1500.

[19] W. Ling, C. Dyer, A. Black, and I. Trancoso, "Two/Too Simple Adaptations of word2vec for Syntax Problems," in *NAACL-HLT*, 2015.

[20] J. P. C. Chiu and E. Nichols, "Named Entity Recognition with Bidirectional LSTM-CNNs," *TACL*, vol. 4, pp. 357–370, 2016.

[21] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF Models for Sequence Tagging," *CoRR*, vol. abs/1508.01991, 2015.

[22] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of Machine Learning Research*, vol. 12, no. Aug, pp. 2493–2537, 2011.

[23] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*. IEEE, 2013, pp. 6645–6649.

[24] N. Reimers and I. Gurevych, "Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging," in *EMNLP*, 2017.

[25] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching Word Vectors with Subword Information," *TACL*, vol. 5, pp. 135–146, 2017.

[26] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proc. of NAACL*, 2018.

[5] `https://biofid.de/en/`

[6] `https://github.com/FID-Biodiversity/GermanWordEmbeddings-NER`

Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)

Deep-EOS: General-Purpose Neural Networks for Sentence Boundary Detection

Stefan Schweter
Bayerische Staatsbibliothek München
Digital Library/Munich Digitization Center
Munich, Germany
{*stefan.schweter*}@*bsb-muenchen.de*

Sajawel Ahmed
Text Technology Lab
Goethe University Frankfurt
Frankfurt, Germany
{*sahmed*}@*em.uni-frankfurt.de*

Abstract

In this paper, we present three general-purpose neural network models for sentence boundary detection. We report on a series of experiments with long short-term memory (LSTM), bidirectional long short-term memory (BiLSTM) and convolutional neural network (CNN) for sentence boundary detection. We show that these neural networks architectures outperform the popular framework of *OpenNLP*, which is based on a maximum entropy model. Hereby, we achieve state-of-the-art results both on multi-lingual benchmarks for 12 different languages and on a *zero-shot* scenario, thus concluding that our trained models can be used for building a robust, language-independent sentence boundary detection system.

1 Introduction

The task of sentence boundary detection is to identify sentences within a text. Many natural language processing (NLP) tasks take a sentence as an input unit, such as part-of-speech tagging (Manning, 2011), dependency parsing (Yu and Vu, 2017), named entity recognition or machine translation. Thus, this foundational task stands at the beginning of various NLP processes and decisively determines their downstream-performance.

Sentence boundary detection is a nontrivial task, because of the ambiguity of the period sign ".", which has several functions (Grefenstette and Tapanainen, 1994), e.g.:

- End of sentence
- Abbreviation
- Acronyms and initialism
- Mathematical numbers

A sentence boundary detection system has to resolve the use of ambiguous punctuation characters to determine if the punctuation character is a true end-of-sentence marker[1].

In the present work, we train different deep architectures of neural networks, such as long short-term memory (LSTM), bidirectional long short-term memory (BiLSTM) and convolutional neural network (CNN), and compare the results with *OpenNLP*[2]. *OpenNLP* is a state-of-the-art tool and uses a maximum entropy model for sentence boundary detection. To test the robustness of our models, we use the *Europarl* corpus for German and English, the *SETimes* corpus for nine different Balkan languages, and the *Leipzig* corpus (Goldhahn et al., 2012) for one Semitic language, namely Arabic. This makes our model language-independent, in which further languages can be used, given the associated training resources are available.

Additionally, we use a *zero-shot* scenario to test our model on unseen abbreviations. We show that our models outperform *OpenNLP* both for each language and on the zero-shot learning task. Therefore, we conclude that our trained models can be used for building a robust, language-independent state-of-the-art sentence boundary detection system.

The remainder of the paper is organized as follows: Section 2 reviews related work. Section 3 presents a sketch of the underlying neural models and the choice of hyperparameters. Section 4 describes the text data and its preprocessing for our twofold experimental setup of a) mono-lingual, and b) zero-shot training. Section 5 reports our results, and, finally, Section 6 discusses our results and draws a conclusion.

2 Related Work

Various approaches have been employed to achieve sentence boundary detection in different languages.

[1] In this paper, we define "?!:;." as potential end-of sentence markers.

[2] *OpenNLP 1.8.4*: `https://opennlp.apache.org`

Recent research in sentence boundary detection focus on machine learning techniques, such as hidden Markov models (Mikheev, 2002), maximum entropy (Reynar and Ratnaparkhi, 1997), conditional random fields (Tomanek et al., 2007), decision tree (Wong et al., 2014) and neural networks (Palmer and Hearst, 1997). Kiss and Strunk (2006) use an unsupervised sentence detection system called *Punkt*, which does not depend on any additional resources. The system use collocation information as evidence from unannotated corpora to detect e.g. abbreviations or ordinal numbers.

The sentence boundary detection task can be treated as a classification problem. Our work is similar to the *SATZ* system, proposed by Palmer and Hearst (1997), which uses a fully-connected feed-forward neural network. The *SATZ* system disambiguates a punctuation mark given a context of *k* surrounding words. This is different to our approach, as we use a char-based context window instead of a word-based context window.

Further high-performers such as *Elephant* (Evang et al., 2013) or *Cutter* (Graën et al., 2018) follow a sequence labeling approach. However, they require a prior language-dependent tokenization of the input text. In contrast to these works, we construct an end-to-end approach which does not depend on the performance of any tokenization method, thus making our *Deep End-Of-Sentence detector* (Deep-EOS) more robust to multi-lingual settings.

3 Model

We use three different architectures of neural networks: long short-term memory (LSTM), bidirectional long short-term memory (BiLSTM) and convolutional neural network (CNN). All three models capture information at the character level. Our models disambiguate potential end-of-sentence markers followed by a whitespace or line break given a context of *k* surrounding characters. The potential end-of-sentence marker is also included in the context window. Table 1 shows an example of a sentence and its extracted contexts: left context, middle context and right context. We also include the whitespace or line break after a potential end-of-sentence marker.

Input sentence	**Left**	**Middle**	**Right**
I go to Mr. Pete Tong	to Mr	.	␣Pete

Table 1: Example for input sentence and extracted context of window size 5.

LSTM We use a standard LSTM (Hochreiter and Schmidhuber, 1997; Gers et al., 2000) network with an embedding size of 128. The number of hidden states is 256. We apply dropout with probability of 0.2 after the hidden layer during training. We apply a sigmoid non-linearity before the prediction layer.

BiLSTM Our bidirectional LSTM network uses an embedding size of 128 and 256 hidden states. We apply dropout with a probability of 0.2 after the hidden layer during training, and we apply a sigmoid non-linearity before the prediction layer.

CNN For the convolutional neural network we use a 1*D* convolution layer with 6 filters and a stride size of 1 (Waibel et al., 1989). The output of the convolution filter is fed through a global max pooling layer and the pooling output is concatenated to represent the context. We apply one 250-dimensional hidden layer with ReLU non-linearity before the prediction layer. We apply dropout with a probability of 0.2 during training.

Other Hyperparameters Our proposed character-based model disambiguates a punctuation mark given a context of *k* surrounding characters. In our experiments we found that a context size of 5 surrounding characters gives the best results. We found that it is very important to include the end-of-sentence marker in the context, as this increases the F1-score of 2%. All models are trained with averaged stochastic gradient descent with a learning rate of 0.001 and mini-batch size of 32. We use Adam for first-order gradient-based optimization. We use binary cross-entropy as loss function. We do not tune hyperparameters for each language. Instead, we tune hyperparameters for one language (English) and use them across languages. Table 2 shows the number of trainable parameters for each model.

Model	**# Parameters**
LSTM	420,097
BiLSTM	814,593
CNN	33,751

Table 2: Number of trainable parameters for LSTM, bidirectional LSTM and CNN.

4 Experimental Setup

Data Similar to Wong et al. (2014) we use the *Europarl* corpus (Koehn, 2005) for our experiments. The *Europarl* parallel corpus is extracted from the proceedings of the European Parliament and is originally created for the research of statistical machine translation systems. We only use German and English from *Europarl*. Wong et al. (2014) does not mention that the *Europarl* corpus is not fully sentence-segmented. The *Europarl* corpus has a one-sentence per line data format. Unfortunately, in some cases one or more sentences appear in a line. Thus, we define the *Europarl* corpus as "quasi"-sentence segmented corpus. We use the *SETimes* corpus (Tyers and Alperen, 2010) as a second corpus for our experiments. The *SETimes* corpus is based on the content published on the *SETimes.com news portal* and contains parallel texts in ten languages. Aside from English the languages contained in the *SETimes* corpus fall into several linguistic groups: Turkic (Turkish), Slavic (Bulgarian, Croatian, Macedonian and Serbian), Hellenic (Greek), Romance (Romanian) and Albanic (Albanian). The *SETimes* corpus is also a "quasi"-sentence segmented corpus. For our experiments we use all the mentioned languages except English, as we use an English corpus from *Europarl*. We do not use any additional data like abbreviation lists. We use the *Leipzig* corpus as the third and final corpus to include the non-European language Arabic into the scope of our investigations. For a *zero-shot* scenario we extracted 80 German abbreviations including their context in a sentence from Wikipedia. These abbreviations do not exist in the German *Europarl* corpus.

Preprocessing All corpora are not tokenized. Text tokenization (or, equivalently, segmentation) is highly non-trivial for many languages (Schütze, 2017). It is problematic even for English as word tokenizers are either manually designed or trained. For our proposed sentence boundary detection system we use a similar idea from Lee et al. (2017). They use a character-based approach without explicit segmentation for neural machine translation. We also use a character-based context window, so no explicit segmentation of input text is necessary.

For all corpora we use the following preprocessing steps: (a) we remove duplicate sentences, (b) we extract only sentences with ends with a potential end-of-sentence marker. Each text collection for a language is split into train, dev and test sets. Table 3 shows a detailed summary of the training, development and test sets used for each language.

Language	# Train	# Dev	# Test
German	1,476,653	184,580	184,580
English	1,474,819	184,352	184,351
Arabic	1,647,906	274,737	276,172
Bulgarian	148,919	18,615	18,614
Bosnian	97,080	12,135	12,134
Greek	159,000	19,875	19,874
Croatian	143,817	17,977	17,976
Macedonian	144,631	18,079	18,078
Romanian	148,924	18,615	18,615
Albanian	159,323	19,915	19,915
Serbian	158,507	19,813	19,812
Turkish	144,585	18,073	18,072

Table 3: Number of sentences in *Europarl*, *SETimes* and *Leipzig* corpus for each language for training, development and test set.

Tasks In the first task we train our different models on the *Europarl*, *SETimes* and *Leipzig* corpus. The second task is to perform *zero-shot* sentence boundary detection. For the *zero-shot* scenario the trained models for the German *Europarl* corpus are used.

Setup We evaluate our different models on our three corpora. We measure F1-score for each model. As baseline to our models, we use *OpenNLP*. *OpenNLP* uses a maximum entropy model. *OpenNLP* comes with pretrained models for German and English, but to ensure a fair comparison between our models and *OpenNLP*, we do not use them. Instead, we train a model from scratch for each language with the recommended hyperparameters from the documentation. For the *zero-shot* scenario we use our trained LSTM, BiLSTM and CNN models on the German *Europarl* corpus and the trained model with *OpenNLP* to perform a zero-shot sentence boundary detection on the crawled abbreviations.

5 Results

We train a maximum of 10 epochs for each model. For the German and English corpus (*Europarl*) the time per epoch is 55 minutes for the BiLSTM model, 28 minutes for the LSTM model and 5 minutes for the CNN model. For each language from the *SETimes* corpus the time per epoch is 5 minutes

Lang.	**LSTM**	**BiLSTM**	**CNN**	***OP***
German	**97.59**	**97.59**	97.50	97.38
English	98.61	**98.62**	98.55	98.40
Arabic	**99.86**	99.83	81.97	99.76
Bulg.	99.22	**99.27**	99.22	98.87
Bosn.	**99.58**	99.52	99.53	99.25
Greek	99.67	**99.70**	99.66	99.25
Croat.	**99.46**	99.44	99.44	99.07
Maced.	98.04	**98.09**	97.94	97.86
Roman.	99.05	99.05	**99.06**	98.89
Alban.	**99.52**	99.51	99.47	99.34
Serbian	98.72	**98.76**	98.73	98.32
Turkish	98.56	**98.58**	98.54	98.08

Table 4: Results on test set for *Europarl*, *SETimes* and *Leipzig* corpus against *OpenNLP* (OP). The highest F1-score for each task on each language is marked in bold face.

for the Bi-LSMT model, 3 minutes for the LSTM model and 20 seconds for the CNN model. Timings are performed on a server machine with a single Nvidia Tesla K20Xm and Intel Xeon E5-2630.

The results on test set on the *SETimes* corpus are shown in Table 4. For each language the best neural network model outperforms *OpenNLP*. On average, the best neural network model is 0.38% better than *OpenNLP*. The worst neural network model also outperforms *OpenNLP* for each language. On average, the worst neural network model is 0.33% better than *OpenNLP*. In half of the cases the bi-directional LSTM model is the best model. In almost all cases the CNN model performs worse than the LSTM and bi-directional LSTM model, but it still achieves better results than the *OpenNLP* model. This suggests that the CNN model still needs more hyperparameter tuning.

The first two rows in Table 4 show the results on test set on the *Europarl* corpus. For both German and English the best neural network model outperforms *OpenNLP*. The CNN model performs worse than the LSTM and bi-directional LSTM model but still achieves better results than *OpenNLP*. The bi-directional LSTM model is the best model and achieves the best results for German and English. On average, the best neural network model is 0.22% better than *OpenNLP*, whereas the worst neural network model is still 0.14% better than *OpenNLP*.

Table 5 shows the results for the *zero-shot* scenario. The CNN model outperforms *OpenNLP* by a large margin and is 6% better than *OpenNLP*. The CNN model also outperforms all other neural network models. Interestingly, the CNN model performs better in a *zero-shot* scenario than in the previous tasks (*Europarl* and *SETimes*). That suggests that the CNN model generalizes better than LSTM or BiLSTM for unseen abbreviations. The worst neural network model (LSTM model) still performs 1,6% better than *OpenNLP*.

Model	**Precision**	**Recall**	**F1**
LSTM	56.62	96.25	71.29
BiLSTM	60.00	97.50	74.29
CNN	61.90	97.50	**75.12**
OpenNLP	54.60	96.25	69.68

Table 5: Results on the *zero-shot* scenario for unseen German abbreviations.

6 Discussion & Conclusion

In this paper, we propose a general-purpose system for sentence boundary detection using different architectures of neural networks. We use the *Europarl*, *SETimes* and *Leipzig* corpus and compare our proposed models with *OpenNLP*. We achieve state-of-the-art results.

The results on the three corpora show that the trained neural network models perform well for all languages. We tune hyperparameters just for one language (English) and share these hyperparameter settings across other languages. This suggests that the proposed neural network models can adopt other languages as well, which makes them language-independent. Our character-based context approach requires no explicit text segmentation and is robust against unknown words.

In a *zero-shot* scenario, in which no manifestation of the test abbreviations is observed during training, our system is also robust against unseen abbreviations. It shows that our proposed neural network models can detect abbreviations "on the fly", after the model has already been trained.

The fact that our proposed neural network models perform well on different languages and on a *zero-shot* scenario leads us to the conclusion that *Deep-EOS* is a *general-purpose* system[3]. Our system can be used for a wide variety of practical use cases, e.g. in the scope of the *BIOfid* project where unstructured OCR text data on biodiversity has to

[3] `https://github.com/stefan-it/deep-eos`

be processed for the task of biological Named Entitiy Recognition (Ahmed and Mehler, 2018; Ahmed et al., 2019).

Acknowledgments

We would like to thank the *Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften* (LRZ) for giving us access to the NVIDIA *DGX-1* supercomputer. Special thanks go to Prof. R. V. Zicari and Prof. A. Mehler for their constructive comments on the final manuscript, to Prof. J. Leidner for his remarks on the related work, and last but not least to Prof. U. Meyer and *Goethe Research Academy for Early Career Researchers* (GRADE) for funding the publication process of this paper.

References

Sajawel Ahmed and Alexander Mehler. 2018. Resource-Size matters: Improving Neural Named Entity Recognition with Optimized Large Corpora. In *Proceedings of the 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*.

Sajawel Ahmed, Manuel Stoeckel, Christine Driller, Adrian Pachzelt, and Alexander Mehler. 2019. BIOfid Dataset: Publishing a German Gold Standard for Named Entity Recognition in Historical Biodiversity Literature. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. Association for Computational Linguistics. accepted.

Kilian Evang, Valerio Basile, Grzegorz Chrupała, and Johan Bos. 2013. Elephant: Sequence labeling for word and sentence segmentation. In *EMNLP 2013*.

Felix A. Gers, Jürgen A. Schmidhuber, and Fred A. Cummins. 2000. Learning to Forget: Continual Prediction with LSTM. *Neural Comput.*, 12(10):2451–2471, October.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *LREC*.

Johannes Graën, Mara Bertamini, and Martin Volk. 2018. Cutter–a Universal Multilingual Tokenizer. In *Proceedings of the 3rd Swiss Text Analytics Conference-SwissText*, pages 75–81.

Gregory Grefenstette and Pasi Tapanainen. 1994. What is a word, What is a sentence? Problems of Tokenization. In *COMPLEX*, pages 79–87.

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

Tibor Kiss and Jan Strunk. 2006. Unsupervised Multilingual Sentence Boundary Detection. *Comput. Linguist.*, 32(4):485–525, December.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Jason Lee, Kyunghyun Cho, and Thomas Hofmann. 2017. Fully character-level neural machine translation without explicit segmentation. *TACL*, 5:365–378.

Christopher D. Manning. 2011. Part-of-speech Tagging from 97% to 100%: Is It Time for Some Linguistics? In *CICLing*, pages 171–189, Berlin, Heidelberg. Springer-Verlag.

Andrei Mikheev. 2002. Periods, Capitalized Words, etc. *Comput. Linguist.*, 28(3):289–318, September.

David D. Palmer and Marti A. Hearst. 1997. Adaptive Multilingual Sentence Boundary Disambiguation. *Comput. Linguist.*, 23(2):241–267, June.

Jeffrey C. Reynar and Adwait Ratnaparkhi. 1997. A Maximum Entropy Approach to Identifying Sentence Boundaries. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, ANLC '97, pages 16–19, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hinrich Schütze. 2017. Nonsymbolic Text Representation. In *EACL*, pages 785–796.

Katrin Tomanek, Joachim Wermter, and Udo Hahn. 2007. Sentence and token splitting based on conditional random fields. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, pages 49–57.

Francis M Tyers and Murat Serdar Alperen. 2010. South-east european times: A parallel corpus of balkan languages. In *Proceedings of the LREC Workshop on Exploitation of Multilingual Resources and Tools for Central and (South-) Eastern European Languages*, pages 49–53.

Alex Waibel, Toshiyuki Hanazawa, Geoffrey Hinton, Kiyohiro Shikano, and Kevin J Lang. 1989. Phoneme recognition using time-delay neural networks. *IEEE transactions on acoustics, speech, and signal processing*, 37(3):328–339.

Derek F Wong, Lidia S Chao, and Xiaodong Zeng. 2014. iSentenizer-μ: Multilingual sentence boundary detection model. *The Scientific World Journal*, 2014.

Xiang Yu and Ngoc Thang Vu. 2017. Character Composition Model with Convolutional Neural Networks for Dependency Parsing on Morphologically Rich Languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 672–678, Vancouver, Canada, July. Association for Computational Linguistics.

BIOfid Dataset: Publishing a German Gold Standard for Named Entity Recognition in Historical Biodiversity Literature

Sajawel Ahmed[1], Manuel Stoeckel[1], Christine Driller[2],
Adrian Pachzelt[3], Alexander Mehler[1]
[1]Goethe University Frankfurt
[2]Senckenberg Nature Research Society
[3]Frankfurt University Library
{sahmed,mehler}@em.uni-frankfurt.de

Abstract

The *Specialized Information Service Biodiversity Research* (*BIOfid*) has been launched to mobilize valuable biological data from printed literature hidden in German libraries for over the past 250 years. In this project, we annotate German texts converted by OCR from historical scientific literature on the biodiversity of plants, birds, moths and butterflies. Our work enables the automatic extraction of biological information previously buried in the mass of papers and volumes. For this purpose, we generated training data for the tasks of *Named Entity Recognition* (NER) and *Taxa Recognition* (TR) in biological documents. We use this data to train a number of leading machine learning tools and create a gold standard for TR in biodiversity literature. More specifically, we perform a practical analysis of our newly generated *BIOfid dataset* through various downstream-task evaluations and establish a new state of the art for TR with 80.23% F-score. In this sense, our paper lays the foundations for future work in the field of information extraction in biology texts.

1 Introduction

Data is the gold to any machine learning (ML). Most ML approaches to *Natural Language Processing* (NLP) address modern, high-resource languages (such as English or Chinese) rather than historical, low-resource languages. As a consequence, feasible ML-tools for processing historical documents are still rare. In this paper we consider corpora of historical German texts in order to extract useful information about biological systems in the past (e.g. species, biotopes etc.).

As a contribution to closing the gap between NLP of modern and of historical languages, we present the newly annotated *BIOfid dataset* for *Named Entity Recognition* (NER) and for *Taxa Recognition* (TR) in the domain of biology, the first of its kind concerning the German language. Our approach is especially designed to address the exploration of biodiversity data[1] from historical documents. We perform a large-scale annotation of scanned texts converted by OCR from historical scientific books on the biodiversity of plants, birds, moths and butterflies, thereby creating the necessary training data to accomplish the task of biological NER and TR using various ML algorithms. Our work facilitates an automatic extraction of biological information so far buried in the bulk of papers and volumes (see Table 1). Overall, our newly generated dataset provides a gold standard and hereby lays the foundations for future work, such as relation extraction and classification based on extracted biological named entities and taxa.

Input sentence:
Ahmed observes that Iris grows in Mai in Frankfurt.
TR output:
Ahmed observes that [Iris]`TAXON` *grows in Mai in Frankfurt.*
Biological NER output:
[Ahmed]`PER` *observes that [Iris]*`TAXON` *grows in [Mai]*`TIME` *in [Frankfurt]*`LOC`*.*

Table 1: Example for our selected tasks.

We perform a practical analysis of our dataset via various downstream-task evaluations. First, we generate a baseline for recognizing taxonomic entities by constructing a sequence tagger based on skip-n-grams and external knowledge resources (i.e. WikiData). Secondly, we apply the best publicly available word embeddings for German and use them alongside our BIOfid dataset as an input for training high-performing neural mod-

[1] Biodiversity is the science which measures the variability and diversity of animals and plants.

Proceedings of the 23rd Conference on Computational Natural Language Learning, pages 871–880
Hong Kong, China, November 3-4, 2019. ©2019 Association for Computational Linguistics

els for NER, namely BiLSTM, ELMo, Flair and BERT (Ahmed and Mehler, 2018; Peters et al., 2018; Akbik et al., 2018; Devlin et al., 2018). By using the optimized BiLSTM model we achieve a new best F-score of 80.23% regarding the recognition of taxonomic entities.

The remainder of the paper is organized as follows: Section 2 reviews related work. Section 3 describes the source texts and the preprocessing pipeline. Section 4 describes the annotation guidelines, process and environment for producing the BIOfid dataset, and methods (n-gram-based sequence tagger, neural models) for evaluating the practical quality of our annotated dataset. Section 5 presents the experimental results. Finally, Section 6 draws a conclusion.

2 Related Work

2018 was a vital year for the task of German NER, following a saturation period from when the last major progress was made by Lample et al. (2016). With the grammar-specific morphological processing and resource-optimization presented by Ahmed and Mehler (2018), the gap between English and German NER was closed. In the same year, with the emergence of multilingual language models such as *ELMo*, *Flair* and *BERT* (Peters et al., 2018; Akbik et al., 2018; Devlin et al., 2018), the performance of various NLP tasks, including NER, was notably improved. Hence, the task of German NER has benefited from these developments.

However, with respect to the availability of a variety of resources, there has not been much progress made until now. Regarding the standard task of NER based on four categories (PERSON, LOCATION, ORGANIZATION, OTHER), the first choice of resources for German is still the *GermEval* dataset (Benikova et al., 2014), followed by the datasets of *CoNLL* and *TüBa-D/Z* (Tjong Kim Sang and De Meulder, 2003; Telljohann et al., 2012). However, their potential for purposes outside of theoretical ML is limited. These datasets do not contain any annotations for taxonomic and temporal entities which are of key interest for biodiversity researchers.

For biological NER in the German language, there are no predecessor resources available to the knowledge of the authors; only an English counterpart exists, namely the *Copious* dataset (T.H. Nguyen et al., 2019), which has been recently published during our ongoing work. This confirms our research endeavors and shows the necessity of more data in this field. We take the English counterpart as the baseline and compare its dataset and results with our own. Overall, our work constitutes the first effort on enabling a state-of-the-art performance for neural representation learning to biological NER.

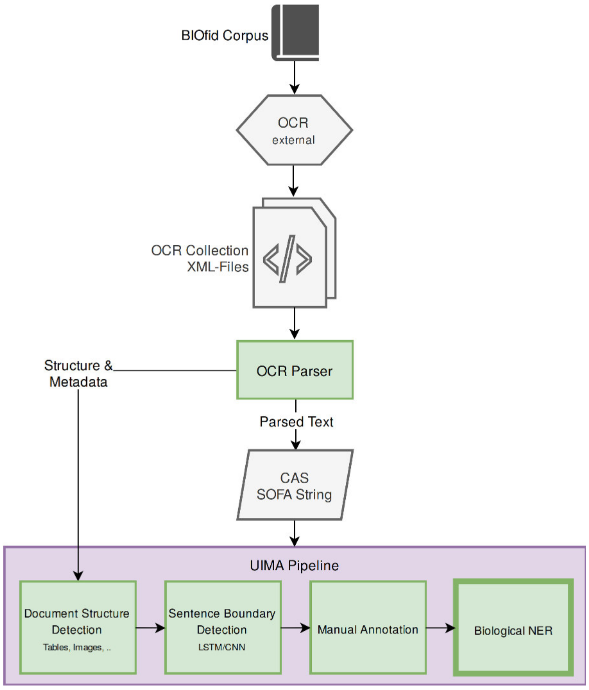


Figure 1: Flowchart showing the data cleaning steps within our preprocessing pipeline.

3 Source Texts & Preprocessing Pipeline

BIOfid Corpus The *BIOfid Corpus* is a collection of historical scientific books on central European biodiversity. It was assembled by a group of German domain experts, denoting a potential pool of relevant print-only journals and publications for historical biodiversity science. However, mainly due to license issues, not all publications could be considered for the corpus.

The available publications were scanned by an external service and subsequently paginated with the software *Visual Library*. Subsequently, every high-resolution page (400 dpi) was digitized with *ABBYY FineReader 8.0 (2005)* to ABBYY-XML, which includes structural information like paragraphs, bold/italic text, images, and table blocks.

OCR Parser The raw OCR data contained various errors, e.g. delivering typical OCR errors such as confusing letters (**ß** → **b**), or delivering gibberish due to the wrong recognition of non-textual elements in scans such as images, figures, or tables. Furthermore, species names or their appended author citation were frequently recognized incorrectly, e.g. ”Lepidium ruderale L.” → ”Lepidium rüderale I.”.

We built the following preprocessing pipeline (see Figure 1) to clean the source data and increase its overall quality. First, the raw OCR data was passed to a parser (labeled ”OCR Parser” in Figure 1). This parser read a given ABBYY-XML into a UIMA CAS, while retaining all structural information in a custom UIMA type system, which was tailored to the ABBYY-XML output.

Using a set of heuristics, the structural information was used to detect erroneous parts in the parsed text, such as page numbers, image and figure blocks mislabeled as text, text margins and table lines parsed as the characters ”I” or ”-”, and tables containing merely non-word characters such as counts of observations[2].

The parser performed further fundamental text segmentation using the information given by the ABBYY-XML, such as tokenization and paragraph splitting. The ABBYY-XML contains tokenization information on the character basis, denoting whether a character is marking the beginning of a word. This information was used alongside plain whitespaces to tokenize the raw text, while further splitting words from non-word characters. All this information was stored in a UIMA CAS using the aforementioned type system and passed down the UIMA pipeline.

Document Structure The BIOfid corpus comprises about 15 journal titles including approximately 410 books. 201 of these books containing 969 articles were selected by domain experts as a representative sample from the entire corpus to generate training data for biological NER.

Sentence Boundary Detection In biological literature, author citations are commonly abbreviated (e.g. Carl von Linné in ”Fagus sylvatica L.”) as well as species names (e.g. ”F. sylvatica” after the first definition). Therefore, standard rule-based tools often fail to detect the correct sentence boundaries in such unstructured raw text documents. Hence, for this task we included the LSTM-based sentence boundary detector *DeepEOS* (Schweter and Ahmed, 2019) in our processing pipeline and trained it with 1,361 sentences, which were manually extracted from the BIOfid corpus. The total amount of training sentences was increased from a preliminary size of 300, since the first experimental results revealed that the SBD is crucial for the performance of our downstream-task.

4 BIOfid Dataset & Methods

4.1 Annotation Guidelines

Named Entities NEs are real-world objects in a given natural language text which denote a unique individual with a *proper name* (e.g. Frankfurt, Africa, Linnaeus, BHL). This stands in contrast to the class of *common names* which refer to some kind of entities (e.g. city, continent, person, corporation) and *not* a uniquely identifiable object.

The standard task of NER focuses on the former class of proper names. However, it is often not easy to differentiate between both classes. Hence, to support the annotators in making the right decision, we created guidelines which demonstrated the rules for annotations. We gradually developed this document in collaboration with the annotators, until finalizing it as the guidelines for annotating the BIOfid corpus. The appendix shows the material which was provided to the team of annotators. First, in Appendix A some introductory examples from the BIOfid corpus are given. Next, in Appendix B the general guidelines used for producing the NER dataset are shown.

As we essentially extend the standard task of NER to our scope of biodiversity, our guidelines are built upon those used for producing the GermEval dataset (Benikova et al., 2014). For this, we take the original German text and extend it with the important adjustments described in the next paragraphs for the context of biodiversity. In contrast to Benikova et al. (2014), we do not consider derivative or partial NEs as a separate category. As the recent work of Ahmed an Mehler (2018) has shown, discarding subtle details is even beneficial, whereas fine-graded feature engineering for deep neural networks usually deteriorates the final performance.

Time In the standard task of NER, temporal information is not captured by the four base entities. However, the aspect of time is important for the research on biodiversity which is constantly evolving. Therefore, we annotated every text unit

[2] An example of such pages is given in Appendix C.

Dataset	Sentence	PERSON	LOCATION	ORGANIZATION	OTHER	TIME	TAXON
CoNLL	18,933	*5,369*	6,579	4,441	3,968	N/A	N/A
GermEval	31,300	10,807	17,275	8,303	4,557	N/A	N/A
TüBa-D/Z	104,787	55,746	28,582	32,224	12,865	N/A	N/A
Copious	*26,277*	*2,889*	*9,921*	*N/A*	*N/A*	*2,210*	*12,227*
BIOfid	**15,833**	**5,393**	**6,785**	**1,085**	**7,849**	**5,197**	**15,085**

Table 2: Statistics for German NER datasets together with the English biological NER dataset *Copious* (T.H. Nguyen et al., 2019).

which denotes a specific temporal entity with the tag TIME (e.g. *[13.02.1835]*TIME, see more in Appendix B: Table 9). For text units which describe a time interval, we marked the starting and ending points as two distinct temporal entities.

Taxonomy Taxonomy is a field in biology that deals with the systematic classification of organisms by morphological, phenotypic, behavioral and phylogenetic characteristics. Based on a variety of common traits, a group of organisms forms a so-called taxon. A well-known example of this are the Darwin's finches, endemic birds in the Galápagos Islands. The different species (each species represents a taxon) are distinguished primarily by the size and shape of their beaks and the associated specialized diets.

Taxa are classified according to international nomenclature codes[3,4,5,6] and are delineated at different hierarchical levels, also known as taxonomic ranks. Most of us are well acquainted with the distinction between the animal and plant kingdoms, although there are other kingdoms e.g. fungi or bacteria. Subordinate to a kingdom are many more ranks such as phylum, class, order, family, genus and species. According to this, the hierarchical classification of the bird species *Struthio camelus*, the common ostrich, from the lowest to the highest taxonomic rank is as follows: *Struthio camelus* (species), *Struthio* (genus), *Struthionidae* (family), *Struthioniformes* (order), *Aves* (class), *Chordata* (phylum), *Animalia* (kingdom). Each scientific name mentioned here along with its taxonomic rank (in parentheses) represents a taxon, meaning a group of organisms with a set of common characteristics being indicative for a common ancestry.

Due to differing and evolving methods of classification, taxonomies are subject to constant change. This also applies to taxonomic nomenclature. Therefore, among others, synonymy and homonymy also play an important role in biology (e.g. there is a plant genus with the name ”Paris”). The relevance of taxonomy for biodiversity research and conservation is fundamental (Thomson et al., 2018), consequently, we considered it justified to introduce the NE-category of TAXON into the process of NER.

For organisms of all taxonomic ranks, we considered scientific names (both accepted and synonyms) and vernacular names, if referring to a certain taxon, as NEs (e.g. *[Struthio camelus]*TAXON or [common ostrich]TAXON, *[Mirza zaza]*TAXON or [northern giant mouse lemur]TAXON, see more in Appendix B: Table 7). Author citation and year, usually appended to the scientific name of a taxon, were tagged as NEs of the categories PERSON and TIME, respectively (e.g. *[Falco]TAXON [Linnaeus]PER [1758]TIME*). Both author and temporal information embedded within the scientific name, were included in the NE TAXON (e.g. *[Carex praecox [Jacq.]PER var. distans]TAXON [Appel]PER*).

4.2 Annotation Process

We performed a single major series of annotations. Instead of just focusing on some inter-agreement value, we performed double checks on existing annotations on given articles through biological experts. This strategy removed the time overload associated to multi-annotations while ensuring a high quality of data.

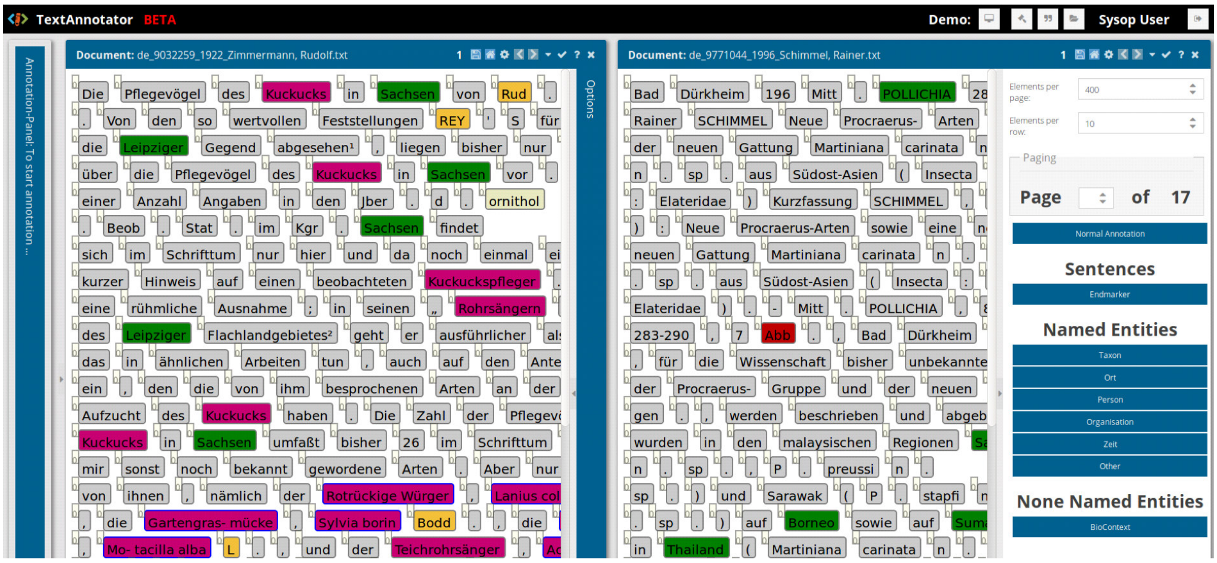
For this scheme, a group of annotators consisting of two researchers from the project team were employed. Both researchers were native speakers of German, and, additionally had a profound background in biology. Besides, two further student assistants with similar profiles were employed to provide further assistance.

[3] http://iczn.org/code

[4] http://www.iapt-taxon.org/nomen/main.php

[5] http://www.the-icsp.org/

[6] http://talk.ictvonline.org/taxonomy/



Document: de_9032259_1922_Zimmermann, Rudolf.txt 1

Die Pflegevögel des Kuckucks in Sachsen von Rud .
. Von den so wertvollen Feststellungen REY ' S für
die Leipziger Gegend abgesehen[1] , liegen bisher nur
über die Pflegevögel des Kuckucks in Sachsen vor .
einer Anzahl Angaben in den Jber . d . ornithol
. Beob . Stat . im Kgr . Sachsen findet
sich im Schrifttum nur hier und da noch einmal ei
kurzer Hinweis auf einen beobachteten Kuckuckspfleger .
eine rühmliche Ausnahme ; in seinen „ Rohrsängern
des Leipziger Flachlandgebietes[2] geht er ausführlicher al:
das in ähnlichen Arbeiten tun , auch auf den Ante
ein , den die von ihm besprochenen Arten an der
Aufzucht des Kuckucks haben . Die Zahl der Pflegev
Kuckucks in Sachsen umfaßt bisher 26 im Schrifttum
mir sonst noch bekannt gewordene Arten . Aber nur
von ihnen , nämlich der Rotrückige Würger , Lanius col
, die Gartengras- mücke , Sylvia borin Bodd . , die
, Mo- tacilla alba L . , und der Teichrohrsänger , Ac

Document: de_9771044_1996_Schimmel, Rainer.txt 1

Bad Dürkheim 196 Mitt . POLLICHIA 28
Rainer SCHIMMEL Neue Procraerus- Arten
der neuen Gattung Martiniana carinata n
n . sp . aus Südost-Asien (Insecta
: Elateridae) Kurzfassung SCHIMMEL ,
) : Neue Procraerus-Arten sowie eine n
neuen Gattung Martiniana carinata n .
. sp . aus Südost-Asien (Insecta :
Elateridae) . - Mitt . POLLICHIA , 8
283-290 , 7 Abb . , Bad Dürkheim
, für die Wissenschaft bisher unbekannte
der Procraerus- Gruppe und der neuen
gen . , werden beschrieben und abgeb
wurden in den malaysischen Regionen Sa
n . sp . , P . preussi n .
sp .) und Sarawak (P . stapfi n
. sp .) auf Borneo sowie auf Suma
in Thailand (Martiniana carinata n .

Elements per page: 400
Elements per row: 10
Page of 17

Figure 2: Working environment for annotating the BIOfid corpus (figure taken from (Abrami et al., 2019)).

4.3 Annotation Environment

We used the *TextAnnotator* (Abrami et al., 2019), a browser-based annotation tool specifically adjusted for this project. Figure 2 shows the working environment which was provided to the annotators. On the left-hand side of the *QuickAnnotator* view, the raw OCR text from the BIOfid corpus is displayed, separated from the choice of annotation tags on the right-hand side. As sentence splitting was part of the annotation task, we did not provide a sentence view. Instead, we provided the whole article, further allowing the annotators to use contextual information while making their decisions.

4.4 Quality of Data

4.4.1 Quantitative Characteristics

Table 2 shows the total amount of annotated sentences along their six NE-categories and compares this with the three major public datasets for German NER. For our BIOfid dataset, we can see the high value of `TIME` and `TAXON` entities which, so far, do not exist for any publicly available dataset.

4.4.2 Data Format

We use the 4-column CoNLL-format which writes each word of a sentence horizontally along its lemma, POS tag and gold label, separating each sentence by an empty new line. For the tagging scheme, we opt for BIO (IOB2). Listing 1 shows an excerpt of the train file in which the entities `TIME`, `PERSON`, `LOCATION`, `TAXON` are marked by our team of annotators for a given sentence from the BIOfid corpus.

Listing 1: Sample sentence from BIOfid dataset

Mein	mein	PPOSAT	O
Sohn	Sohn	NN	O
konnte	können	VMFIN	O
am	an	APPRART	O
3	3	CARD	B–TME
.	—	$.	I –TME
1	1	CARD	I –TME
.	—	$.	I –TME
23	23	CARD	I –TME
den	der	ART	O
Fabrikanten	Fabrikant	NN	O
Walter	Walter	NE	B–PER
Schmidt	Schmidt	NE	I –PER
aus	aus	APPR	O
Geithain	Geithain	NE	B–LOC
bei	bei	APPR	O
einem	ein	ART	O
Spaziergang	Spaziergang	NN	O
auf	auf	APPR	O
dem	der	ART	O
Rochlitzer	Rochlitzer	NN	B–LOC
Berge	Berg	NN	I –LOC
auf	auf	APPR	O
eine	ein	ART	O
Ringamsel	Ringamsel	NN	B–TAX
,	—	$,	O
Turdus	Turdus	NN	B–TAX
torquatus	torquatus	ADJD	I –TAX
L	L	NN	B–PER
.	—	$.	O
,	—	$,	O
hinweisen	hinweisen	VVINF	O
.	$.	—	O

We split the BIOfid dataset into train, dev, test files by the common ratio of 80:10:10 percentages after randomizing its order of sentences. These final data files are utilized for training and evaluating our models, which are described in the next section.

4.5 Methods

For the evaluation of the BIOfid dataset, we use six different approaches and compare each others results: one classic *rule-based* model and five high-performing *embedding-based* models.

4.5.1 N-Gram Tagger for TR

We develop a naive sequence tagger as a baseline for the recognition of taxonomic entities in the BIOfid dataset. The baseline is only for a subtask of the full task of biological NER, described in the previous Section 4.1. Our sequence tagger is built on the k-skip-n-grams (with $k = 1$) which are constructed from the tokens of taxonomic entries in the comprehensive *Latin* and *German* gazetteers of biology. Both gazetteers consist of 83,348 taxonomic entries from various biological systematics such as of *aves*, *lepidoptera* and *vascular plant*. In addition, we consider *WikiData*[7] and construct an additional gazetteer by extracting 2,663,995 German and Latin taxonomic entries from the online resource by selecting all entries from a XML-dump that are subjects (`?s`) in the following two SPARQL triple patterns[8]:

- `?s instance-of taxon.`
- `?o subclass-of taxon.`
  `?s instance-of ?o.`

For each gazetteer entry consisting of at least three tokens ($n \geq 3$), we take all tokens as an input and create a list of 1-skip-n-grams. For example, for the taxonomic entry *iris kashmiriana b.*, we create four n-grams *(iris kashmiriana), (iris b.), (kashmiriana b.)* and *(iris kashmiriana b.)*. In this way, we construct 3,023,270 unique n-grams in total from 2,682,959 merged taxonomic entries, while dropping 140,432 duplicate n-grams entirely. Next, we map all these n-grams to the BIOfid test file by standard string matching and thus find the taxonomic occurrences in the target set of text data.

4.5.2 Neural Models for NER

Our neural models consist of two separately trained components: a) foundational word embeddings, modeling the general knowledge from large unlabeled text corpora, and b) various task-specific neural architectures, modeling the domain knowledge from the labeled training data. In this section, both components are presented briefly.

Word Embeddings The language model of continuous space word representations (*word2vec*) (Mikolov et al., 2013) and its variations by (Levy and Goldberg, 2014; Komninos and Manandhar, 2016) are the foundations of most ongoing research in NLP with neural networks. Based on the context, the model embeds words, phrases or sentences into high dimensional vector spaces. We use the model of *Wang2vec* (Ling et al., 2015) and its morphological extension (Ahmed and Mehler, 2018) which explores syntactic data specific for German and, thus, better suites the task of NER. We use the recently published German language word embeddings from the TTLab[9] which are pre-trained with the morphological extension of the Wang2vec algorithm on the COW corpus (Schäfer, 2015), the largest collection of German texts extracted from web documents with over 617 Mio. sentences. Out of the six published variants of embeddings, we opt for token-based embeddings (*COW.lower.wang2vec*), as they delivered the best results for German NER according to the publishers.

BiLSTM We provide a brief overview of the configurations for the five neural models which we use throughout this paper. The model *BiLSTM-CRF* is similar to the one used in (Ahmed and Mehler, 2018), which goes back to the work of (Lample et al., 2016). The neural network consists of stacked LSTM and CRF layers. The *base layer* combines for a given word its (pre-trained) word embedding with its character-based embedding. These features are forwarded to the *prediction layer* which produces the final NE tag.

Model	Emb.	Language Model	Train Data
BiLSTM-a	COW	N/A	BIOfid
Flair Wang2v.	COW	PCE	BIOfid
Flair ELMo	COW	PCE+Leipzig	BIOfid
Flair BERT	COW	PCE+BERT-Base	BIOfid
BiLSTM-b	COW	N/A	All

Table 3: Overview of the model inputs. For BiLSTM-b we consider all merged training data (i.e. BIOfid + GermEval + CoNLL)

Flair Wang2vec We further train a sequence labeling model using Flair[10]. We build the model in

[7] http://www.wikidata.org/

[8] All results can be acquired with the following WikiData queries: http://w.wiki/3u3 and http://w.wiki/3ud

[9] http://www.texttechnologylab.org/resources2018/

[10] http://github.com/zalandoresearch/flair

the same fashion as used by (Akbik et al., 2018) following the guide given by the authors for the task ”CoNLL-03 Named Entity Recognition (German)”, while keeping the pooled contextualized embeddings (PCE) and exchanging the GloVe embeddings employed by the authors with Wang2vec embeddings trained on the COW corpus.

Flair ELMo In addition to the previous model, we train a Flair Sequence Tagging model by stacking an ELMo embedding layer on top of the Flair Wang2vec model. The ELMo embeddings were trained on a section of the *Leipzig Corpora Collection* (Goldhahn et al., 2012) containing 100,000 sentences from Wikipedia using default parameters.

Flair BERT Similarly, we added BERT (Devlin et al., 2018) to the Flair Wang2vec model. We used the recently published *BERT-Base, Multilingual Cased*[11] pre-trained model for this purpose.

Hyperparameters We take the original neural models and keep the hyperparameters as described in their references. The only adjustments we make to the models are on the input level, i.e. we perform variations for the pre-trained word embeddings, the pre-trained language models, and the training data (see Table 3).

5 Results

We evaluate the performance of all models with the official script from the shared task of CoNLL 2003 (Tjong Kim Sang and De Meulder, 2003). All our experiments were run on Nvidia's *GTX 1080 Ti* GPUs.

5.1 Baseline for TR

N-Gram Tagger Applying the gazetteer to the BIOfid test file gives us the respective baseline for the recognition of taxonomic entities. For evaluation, we use the CoNLL-script and contrast it with easing the conditions by evaluating only the NE predictions and ignoring the prefixed BIO-tagging scheme to every NE. The evaluation does not take into account the other words and is based only on the actual words annotated as TAXON.

Table 4 displays the results for the n-gram tagger. We can nicely see that the increase in size of gazetteers leads to an increase in the final performance. More specifically, for the eased condition, every incremental step from *Latin* to *Latin+German*, and the next step to *All* (i.e. *Latin+German+WikiData*) leads to an increase of +6.64% and +6.56% F-scores, respectively. This matter of fact demonstrates that for the n-gram tagger the resource-size matters.

Gazetteer	CoNLL-Eval	Pr. [%]	Re. [%]	F1 [%]
Lat.	standard	61.50	34.71	44.37
Lat.+Ger.	standard	65.83	45.42	53.75
WikiData	standard	69.05	53.91	60.55
Lat.	eased	92.48	46.04	61.06
Lat.+Ger.	eased	92.94	54.55	67.70
WikiData	eased	95.55	58.87	72.85
All	standard	69.20	55.75	61.75
All	eased	95.57	60.72	74.26

Table 4: Baseline for TR on the BIOfid test file with the N-Gram sequence tagger.

Furthermore, for the eased condition, we see very high scores for precision, however, the recall values are relatively low. This result demonstrates a classic problem of rule-based approaches; as there is no learning process involved, we assume that the performance of the n-gram tagger is highly limited on the features extracted from the source of knowledge (i.e. the amount of information contained in the gazetteer). Besides, no transfer learning is possible from related resources, demonstrating the downsides of non-learning methods.

5.2 Biological NER

We report here the results of our comprehensive survey of five current embedding-based high-performers for biological NER in historical biodiversity literature[12].

The Gold Standard Table 5 contains a detailed summary of all results. In that table, we the report the results which are given by T.H. Nguyen et al. (2019). For the optimized *BiLSTM Tagger*, we achieve excellent results and establish a new state-of-the-art for the first task of TR with 80.23% F-score (see Table 5: BiLSTM-a). For biological NER, we outperform the English counterpart *Co-*

[11] http://github.com/google-research/bert

[12] Our manual inspection of the training data showed that the annotations are content-wise homogeneous, except for the category OTHER. The annotators reported its usage as a residual NE-category for everything which is biologically interesting (e.g. morphology, animal behavior, reproduction, development) but does not fall under the definition of the five major categories. Initial experimental results confirmed its heterogeneous quality. Therefore we omitted OTHER (3,143 sentences) from our further experiments which in turn increased the final performance of NER.

Model	Scores [%]	TAXON	PERSON	LOCATION	ORGANIZATION	TIME	*Overall*
Copious Nguyen (2019)	Precision	77.42	58.92	85.05	N/A	70.67	77.49
	Recall	69.67	48.44	85.63	N/A	54.36	71.89
	F1	73.34	53.17	85.34	N/A	61.45	74.58
BiLSTM-a	Precision	81.33	63.19	66.20	60.24	91.16	75.62
	Recall	79.16	77.45	57.35	67.57	88.16	74.98
	F1	**80.23**	69.60	61.46	63.69	89.63	75.30
Flair Wang2vec	Precision	75.94	61.25	67.58	61.64	90.59	73.58
	Recall	81.37	76.09	62.89	58.11	85.24	75.89
	F1	78.08	71.89	62.63	56.95	87.89	74.30
Flair ELMo	Precision	75.64	67.16	58.31	56.82	90.49	73.05
	Recall	79.92	79.89	65.06	60.81	86.02	76.50
	F1	77.88	69.34	66.30	61.22	88.25	75.01
Flair BERT	Precision	76.63	65.30	66.96	58.00	92.21	74.98
	Recall	77.38	81.02	61.89	58.00	90.33	76.22
	F1	77.01	72.31	64.32	58.00	**91.26**	75.59
BiLSTM-b	Precision	80.45	88.61	72.72	81.21	87.63	79.35
	Recall	76.65	89.40	84.02	70.74	81.17	75.38
	F1	78.50	**89.00**	**77.96**	**75.61**	84.27	**77.31**

Table 5: Results for the task of German biological NER with various neural networks models along the English baseline on the Copious dataset (T.H. Nguyen et al., 2019). All models are trained on the BIOfid dataset and evaluated with the official CoNLL-2003 eval script.

pious for all categories except for LOCATION. For the latter category, the *Copious* dataset contains *9,921* training samples whereas ours has 3,136 fewer samples. We assume that this lower amount results into the lower performance.

With the popular deep language models *Flair*, *ELMo* and *BERT*, we interestingly stay below the performance of the BiLSTM model (except for TIME). Although we utilize the same pre-trained COW word embeddings for all models, we assume that the lower performance arises due to the language models themselves being trained on only a relatively small corpus (ELMo: 100,000 sentences). However, for training ELMo on larger corpora, such as the COW corpus, we would require many months of training time. For the pre-trained Flair and BERT, we can only fine-tune the last tagging layer, not the whole language model itself. This stands in contrast to the BiLSTM model which can be wholly targeted to our domain-specific training data. Hence, this demonstrates the downside of such heavy language models; although they might deliver the top performances, it is difficult to adjust them for lightweight processes, making them impractical for the context of low-resources scenarios.

Data Merging for BiLSTM Tagger For BiLSTM-a, it can be noted that the performance of the standard categories PERSON, ORGANIZATION, and, especially LOCATION is inferior. Therefore, we performed resource-optimization by merging high quality data with our BIOfid dataset in order to increase the training samples for the low performing categories. We merge the datasets of GermEval and CoNLL with our annotated sentences, resulting in train, dev, and test sizes of *46,857*, *6,629*, and *9,437* sentences, respectively. Table 5: BiLSTM-b shows the improvements in performance with the increased dataset. Our results demonstrate the effectiveness of our approach; we do not need to modify the model, rather it is sufficient to perform data-driven optimization. Considering the overall performance, we outperform the English counterpart by $+2,73\%$ F-Score and thus establish a new state-of-the-art for the task of biological NER.

Error Analysis We manually analyze the errors made by the ensemble of neural models. We observe three major issues that compose the absolute majority of errors: a number of *missing annotations* from our experts, *OCR erros* in the raw text and *rare words* that occur frequently in our test dataset. An example of an OCR error is the annotated text span *[1, Juni 1967]* TIME which is misclassified by all models as *1, [Juni 1967]* TIME due to the comma in the date format. Another example is *[KLeebend]* LOC which is not tagged due to the capital ”L”. Further, the word *[Venn-*

fußfläche] LOC occurs 17 times in the test dataset, but only twice in the training set. It is a three word compound of the words *Venn*, *Fuß* and *Fläche*, that describes a part of the landscape *Vennvorland* in Germany. We conclude that the preprocessing pipeline has to be further refined to remove the OCR errors, while a re-annotation of the data could solve the missing annotations and a more thorough shuffle may solve the rare word issue.

6 Conclusion

In this study, we presented a newly annotated *BIOfid* dataset for German NER in historical biodiversity literature and performed a comprehensive evaluation of the quality of our dataset with five competing neural models. We come to the conclusion that the value of our dataset does not rely solely on the two new entities of TIME and TAXON. By generating domain-related annotation data typical for historical biodiversity literature, we increase the potential performance for biological NER, even for the four standard NE categories. This was demonstrated by the limited scope of the rule-based approach which could not come close to the performance delivered by the neural models and which, in turn, established a new state-of-the-art for both of our selected tasks of TR and NER.

In the course of the annotation process, we discovered that there are further information entities in the BIOfid corpus which do not fall into the definition of standard NE-categories, albeit they are useful from the perspective of biodiversity researchers. For future work, we plan to increase the semantic granularity of the BIOfid dataset by mapping and re-annotating the existing six NE-categories to the top-level hierarchy of *WordNet* (Miller, 1995). This includes 26 categories that can be either *abstract entities* or *concrete entities* (i.e. NE) and can be assigned to specific biological entities, such as morphology, habitat, reproduction, behavioral traits, or species community. By re-annotating the dataset we additionally plan to deliver an inter-agreement value for both the current NER-dataset and the much smaller WordNet-dataset (which is planned to contain an up to 9 times higher amount of annotated information per sentence). Furthermore, we plan to extract all biological entities with the trained neural models from the BIOfid corpus and perform on them the task of *relation extraction* based on current embedding methods.

Overall, our work mobilizes data from undigitized literature leading to huge potentials for biodiversity researchers. It enables cartographic research on the distribution of Central European biodiversity ranging from the pre-modern time up to our current ever increasingly digitizing age.

Acknowledgments

This work was funded by the German Research Foundations (DFG) as part of the *BIOfid* project[13] (DFG-326061700). The BIOfid dataset, the taxonomic gazetteers, the source code, and the supporting appendix (which includes the full annotation guidelines) are uploaded on GitHub[14] for the research community. Special thanks go to W. A. Hemati and G. Abrami for setting up the annotation environment, and to R. Roller, S. Löbbert, Dr. G. Kasperek, Dr. A. Hausinger, Dr. T. Hörnschemeyer and further project members for supporting the annotation process.

References

Giuseppe Abrami, Alexander Mehler, Andy Lücking, Elias Rieb, and Philipp Helfrich. 2019. TextAnnotator: A flexible framework for semantic annotations. In *Proceedings of the Fifteenth Joint ACL - ISO Workshop on Interoperable Semantic Annotation, (ISA-15)*, ISA-15.

Sajawel Ahmed and Alexander Mehler. 2018. Resource-Size matters: Improving Neural Named Entity Recognition with Optimized Large Corpora. In *Proceedings of the 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual String Embeddings for Sequence Labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Darina Benikova, Christian Biemann, and Marc Reznicek. 2014. NoSta-D Named Entity Annotation for German: Guidelines and Dataset. In *LREC*.

Armin Burkhardt. 2004. 2004. Nomen est omen? : zur Semantik der Eigennamen. In *Landesheimatbund Sachsen-Anhalt e. V. (Hrsg.): "Magdeburger Namenlandschaft" : Orts- und Personennamen der Stadt und Region Magdeburg*.

Hai Leong Chieu and Hwee Tou Ng. 2002. Named Entity Recognition: A Maximum Entropy Approach

[13] http://biofid.de/en/

[14] https://github.com/texttechnologylab/BIOfid

Using Global Information. In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1*, COLING '02, pages 1–7. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *LREC*.

Alexandros Komninos and Suresh Manandhar. 2016. Dependency Based Embeddings for Sentence Classification Tasks. In *HLT-NAACL*, pages 1490–1500.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural Architectures for Named Entity Recognition. In *NAACL-HLT*.

Omer Levy and Yoav Goldberg. 2014. Dependency-Based Word Embeddings. In *ACL (2)*, pages 302–308.

Wang Ling, Chris Dyer, Alan Black, and Isabel Trancoso. 2015. Two/Too Simple Adaptations of word2vec for Syntax Problems. In *NAACL-HLT*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

George A Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Roland Schäfer. 2015. Processing and querying large web corpora with the COW14 architecture. In *Proceedings of Challenges in the Management of Large Corpora 3 (CMLC-3)*, Lancaster. UCREL, IDS.

A Schiller, S Teufel, C Stöckert, and C Thielen. 1999. Guidelines für das Tagging deutscher Textcorpora mit STTS [Guidelines for tagging German corpora of written language with STTS]. Technical report, Technical Report. Stuttgart, Germany: Institut für maschinelle Sprachverarbeitung [Institute for Machine Language Processing].

Stefan Schweter and Sajawel Ahmed. 2019. Deep-EOS: General-Purpose Neural Networks for Sentence Boundary Detection. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS)*. Accepted.

Heike Telljohann, Erhard W Hinrichs, Sandra Kübler, Heike Zinsmeister, and Kathrin Beck. 2012. Stylebook for the Tübingen treebank of written German (TüBa-D/Z).

Scott A Thomson, Richard L Pyle, Shane T Ahyong, Miguel Alonso-Zarazaga, Joe Ammirati, Juan Francisco Araya, John S Ascher, Tracy Lynn Audisio, Valter M Azevedo-Santos, Nicolas Bailly, et al. 2018. Taxonomy based on science is necessary for global conservation. *PLoS biology*, 16(3):e2005075.

Nhung T.H. Nguyen, Roselyn S. Gabud, and Sophia Ananiadou. 2019. COPIOUS: A gold standard corpus of named entities towards extracting species occurrence from biodiversity literature. *Biodiversity Data Journal*, 7:e29626.

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 the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 142–147. Association for Computational Linguistics.

A Examples for annotating named entities in BIOfid corpus

1. *[Coeloglossum viride]*`TAXON` *blüht im [Mai]*`TIME` *auf den Wiesen besonders der [Grabenwiese]*`LOC` *vor dem [Eschenheimer Tor]*`LOC`*, wo ich sie seit [1729]*`TIME` *im [Mai]*`TIME` *fand gross und klein mit ganz grünen Blumen, auch mit einem dunkelroten Bart, mit breiten und schmalen Blättern. Einige haben auch einen Geruch, andere nicht. Zwischen [Falkenstein]*`LOC` *und [Cronberg]*`LOC` *auf Wiesen.*

2. *Das folgende stellt einen Versuch dar, aus dem Verlauf der [nacheiszeitlichen]*`OTH` *Ausbreitung der [Weissbuche]*`TAXON` *oder [Hainbuche]*`TAXON` *([Carpinus betulus]*`TAXON`*) in [Norddeutschland]*`LOC` *einen Beitrag für die Beurteilung des Klimas namentlich zur [späten Wärmezeit]*`OTH` *zu gewinnen.*

3. *Während der Ausgrabung einer grösseren [bandkeramischen Siedlung]*`OTH` *bei [Bracht]*`LOC` *nördlich [Marburg]*`LOC` *wurde Herr Dr. [O. UENZE]*`PER`*, [Amt für Bodenaltertümer]*`ORG`*, [Marburg]*`LOC`*, auf ein der Fundstelle unmittelbar benachbartes kleines Moor von etwa 50m Durchmesser aufmerksam und vermutete, dass es die Wasserstelle der [Neolithiker]*`OTH` *gewesen sei.*

4. *[Falco]*`TAXON` *[Linnaeus]*`PER` *[1758]*`TIME`

5. *[Carex praecox [Jacq.]*`PER` *var. distans]*`TAXON` *[Appel]*`PER`

6. *Verfasser untersuchte die Winterknospen von [S. lanata]*`TAXON`*, [glauca]*`TAXON`*, [lapponum]*`TAXON`*, [phylicifolia]*`TAXON`*, . . .*

7. *[Zug von [Falco vespertinus]*`TAXON` *durch [Westeuropa]*`LOC` *im [September 1927]*`TIME.]OTH` *[Ornithol. Monatsber.]*`ORG` *36 ([1928]*`TIME`*) S.42-44.*

B Annotation Guidelines (biologized version of Benikova et al. (2014))

Guidelines für die Named Entity Recognition. Sie bauen auf den Guidelines in den STTS-Guidelines (Schiller et al., 1999), (Telljohann et al., 2012) und (Chieu and Ng, 2002) auf.

B.1 Einführung: Named Entity Recognition

Unter der Named Entity Recognition (NER) versteht man die Aufgabe, Eigennamen (named entities) in Texten zu erkennen. Technisch gesehen sind hierzu zwei Schritte notwendig. Zuerst müssen in einem laufenden Text die Token gefunden werden, die zu einem Eigennamen gehören (Named Entity Detection: NED), danach können diese Eigennamen semantischen Kategorien zugeordnet werden (Named Entity Classification). Prototypisch ist dabei der Unterschied zwischen Eigennamen und Appellativa der, dass letztere eine Gattung oder eine Klasse beschreiben, während erstere einzelne Individuen oder Sammlungen von Individuen unabhängig von gemeinsamen Eigenschaften bezeichnen (Burkhardt, 2004). Die vorliegenden Guidelines sollen es Annotatoren ermöglichen, Eigennamen in Texte aus Standard und Nichtstandard-Varietäten konsistent zu annotieren. In diesen Guidelines werden die beiden Aufgaben der NED und NEC nicht unterschieden, da die Konzentration auf Beispiele in diesem Dokument, die Trennung künstlich erzeugen müsste und nicht zu erwarten ist, dass die Resultate sich dadurch verbessern würden. In Anlehnung an die oben genannten Guidelines für Zeitungssprache werden in NoSta-D-BIOfid sechs semantische Hauptklassen unterschieden (Personen, Taxa, Organisationen, Orte, Zeiten und Andere).

B.2 Wie finde ich eine NE?

Schritt 1: Nur volle Nominalphrasen können NEs sein. Pronomen und alle anderen Phrasen können ignoriert werden.
Schritt 2: Namen sind im Prinzip Bezeichnungen für einzigartige Einheiten, die nicht über gemeinsame Eigenschaften beschrieben werden.
Beispiel:
[Der Struppi] folgt [seinem Herrchen].
Hier gibt es zwei Nominalphrasen als Kandidaten für einen Eigennamen (NE). “Der Struppi” bezeichnet eine einzige Einheit. Es kann auch mehrere Struppis geben, aber diese haben an sich keine gemeinsamen Eigenschaften, bis auf den gemeinsamen Namen, daher handelt es sich um einen Eigennamen. *”seinem Herrchen”* bezeichnet zwar (typischerweise) auch nur eine einzige Person allerdings können wir diese nur über die Eigenschaft identifizieren, dass sie ein Herrchen ist und dass dies für Struppi zutrifft. Struppi

könnte auch mehrere Herrchen haben, die alle die Eigenschaften teilen, die ein Struppi-Herrchen beinhaltet (z.B. darf Struppi streicheln, muss ihn ausführen und füttern etc.)

Schritt 3: Determinierer sind keine Teile des Namens.
Beispiel: *Der [Struppi]*`NE` *folgt seinem Herrchen.*

Schritt 4: Eigennamen können mehr als ein Token beinhalten. Beispiel:
Viele Personennamen (PER für person):
[Carl Linnaeus]`PER`
Buchtitle (OTH für other):
[Systema Naturae]`OTH`

Schritt 5: Eigennamen können auch in einander verschachtelt sein. Beispiel:
Personennamen in Filmtiteln:
[[Shakespeare]`PER` *in Love]*`OTH`
Orte (LOC für location) in Vereinsnamen (ORG für organisation):
[Hebarium Senckenbergianum [Frankfurt]`LOC`*]*`ORG`

Schritt 6: Titel, Anreden und Besitzer gehören NICHT zu einem komplexen Eigennamen. Besitzer können natürlich selber Eigennamen sein. Beispiel:
Referenz auf Musiktitel:
[Vivaldis]`PER` *[Vier Jahreszeiten]*`OTH`
Referenz auf Personen:
Landesvorsitzende Frau Vorstandsvorsitzende Dr. [Ute Wedemeier]`PER`

Schritt 7: Wenn das Gesamttoken einen Eigennamen darstellt, dann wird dieser annotiert. Beispiel:
Stiftungen: *[[Böll]*`PER`*-Stiftung]*`ORG`

Schritt 8: Kann in einem Kontext nicht entschieden werden, ob eine NP sich als Eigennamen oder Appellativ verhält, wird es nicht als NE markiert. Beispiel:
Ortsnamen vs. -beschreibungen:
...und zogen mit ihren grossen Transparenten gestern vom [Steintor] über den [Ostertorsteinweg]`LOC` *zum [Marktplatz].*

Schritt 9: Wenn ein Name als Bezeichnung für bestimmte Gegenstände in die Sprache übergegangen ist und in seiner Nutzung nicht als NE fungiert, so wird dieser nicht annotiert. Beispiel:
[Teddybär] (`NICHT PER`*)*
[Colt] (`NICHT PER`*)*

Schritt 10: Bei Aufzählungen mit Hilfe von Bindestrichen oder Vertragen eines Teils der NE auf spätere Wörter, wird die NE so annotiert, als sei sie voll ausgeschrieben.
Beispiel:
[Frühe]`OTH` *und [Späte Bronzezeit]*`OTH`
[Süd-]`LOC` *und [Nordafrika]*`LOC`

B.3 Zu welcher semantischen Klasse gehört ein Eigenname?

Wenn der Eigenname in eine der Klassen in der Liste Faustregel zur Unterscheidung einer Klassenbezeichnung und eines Namens gehört, dann annotiere die zugehörige Klasse. Sollte die gefundene NE Rechtschreibfehler enthalten, wird sie dennoch annotiert. In Zweifelsfällen hilft auch die Tabelle NoSta-D-BIOfid-TagSet und alle Untertabellen, insbesondere die Beispiele mit dem weiter.
Jahreszahlen in ORGanisationen werden nicht markiert.
Beispiel:
[ICEI]`ORG` *[2018]*`TIME`
[Fussball-WM]`ORG` *[2014]*`TIME`
Wenn der Eigennamen in KEINE der vorhandenen Klassen passt, markiere diesen mit ***UNCLEAR***, notiere dir bitte das Beispiel und schicke uns eine E-Mail an: `a.b@c.de`. So können wir die Guidelines sukzessiv verbessern.

B.4 Faustregel zur Unterscheidung einer Klassenbezeichnung und eines Namens:

- Elemente der fraglichen Einheit verbinden die gleichen Eigenschaften → Klasse → keine NE

- Christen glauben an Christus → Christ glaubt an Christus → keine NE

- Die Elemente der fraglichen Einheit verbindet nur der Name oder Element ist Einheit bezeichnet ein spezifisches Individuum → Name → NE

- *"Paleocene"* bezeichnet spezifische Epoche → NE (OTH)

NoSta-D-BIOfid-Tagset

Subcategory	**Examples**
person	*Carl Linnaeus*
Surname	*Tüxen, Tx.*
Artist names	*Madonna*
Charactere	*Schneewitchen, Miss Piggy*
Nicknames	*Sternchen333*
Superheroes	*Batman*

Table 6: Category 'PER-Person'

Subcategory	**Examples**
Hybrids	*Abies alba x Abies normannia*
Variety	*Asplenium scolopendrium var. crispum*
Form	*Araschnia levana f. prorsa*
Subspecies	*Falco peregrinus subsp. calidus, Pollichia semirubella ssp. semirubella*
Species	*Coeloglossum viride, Grüne Hohlzunge*
Genus	*Dendrocopus, Buntspechte*
Subfamily	*Phyticinae*
Family	*Noctuidae, Rosaceae*
Order	*Lepidoptera*
Class	*Aves, Insecta*
Phylum	*Chordata, Tracheophyta*
Kingdom	*Animalia, Plantae*

Table 7: Category 'TAXON': scientific and vernacular names (vernacular names only when referring to a certain taxon)

Subcategory	**Examples**
Districts	*Schöneberg*
Sights, Churches	*Brandenburger Tor, Johanniskirche*
Planets	*Mars*
Landscapes	*Königsheide*
Streets, places	*Söogestrasse, Alexanderplatz, A5*
Shopping centres	*Luisencenter, Allee-Center*
Mountains, lakes, rivers	*Alpen, Viktoriasee, Spree*
Continents	*Europa, Asien*
Countries, states	*Frankreich, Hessen, Assyrien, USA*
Cities	*Berlin, Babylon*
Regions	*Gazastreifen*

Table 8: Category 'LOC-Location'

Subcategory	**Examples**
Day	*Freitag*
Month	*Februar*
Year	*1835*
dd.mm.yyyy	*13.02.1835*
Century	*19. Jahrhundert*

Table 9: Category 'TIME'

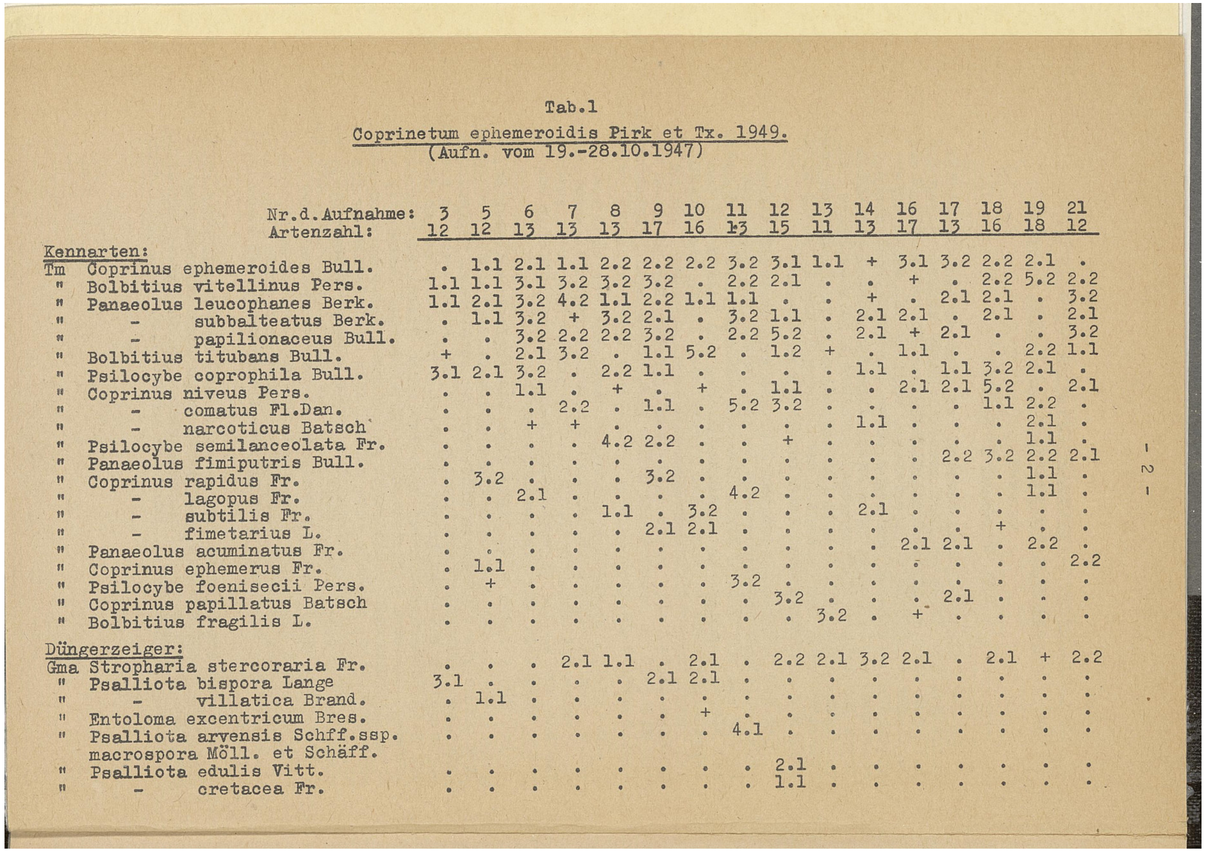
Subcategory	**Examples**
Book-, Film titles etc.	*Faust, Canon Medicinae*
Currencies	*Euro, Deutsche Mark*
Languages	*Deutsch, Latein*
Epochs	*Paleocene, Neolithikum, (auch Neubildungen: 'Neuzeit')*

Table 10: Category 'OTH-Others'

Subcategory	**Examples**
Organisations	*BHL, EU, Landgericht Frankfurt, Deutsche Botanische Gesellschaft*
Companies	*Microsoft, Bertelsmann*
Airports	*Fraport*
Operators	*Lotto 6 aus 49*
Institute	*Institut für Informatik*
Museums	*Senckenberg Museum*
Newspapers, journals	*Süddeutsche Zeitung, Nature, Beiträge zur Entomologie*
Clubs	*Eintracht Frankfurt*
Theatres, cinemas	*Metropol-Theater, CinemaxX*
Festivals	*Eurovision Song Contest, Berlinale*
Expositions	*Faszination Vielfalt*
Universities	*Goethe Universität Frankfurt*
Radio stations	*Arte, Planet Radio*
Restaurants and hotels	*Sassella, Mariott*
Military units	*Blauhelme*
Hospitals, Nursing home	*Charit, Klinikum Ingolstadt*
Fashion brands	*Chanel*
Sporting events	*Olympische Spiele, Wimbledon*
Bands	*Beatles, Die Fantastischen Vier*
Institutions	*DFG, Vogelwarte Helgoland*
Libraries	*UB J.C. Senckenberg*
Parties	*SPD, CDU*

Table 11: Category 'ORG-Organisation'

C Sample Page



Tab.1

Coprinetum ephemeroidis Pirk et Tx. 1949.
(Aufn. vom 19.-28.10.1947)

	Nr.d.Aufnahme:	3	5	6	7	8	9	10	11	12	13	14	16	17	18	19	21
	Artenzahl:	12	12	13	13	13	17	16	13	15	11	13	17	13	16	18	12
Kennarten:																	
Tm	Coprinus ephemeroides Bull.	.	1.1	2.1	1.1	2.2	2.2	2.2	3.2	3.1	1.1	+	3.1	3.2	2.2	2.1	.
"	Bolbitius vitellinus Pers.	1.1	1.1	3.1	3.2	3.2	3.2	.	2.2	2.1	.	.	+	.	2.2	5.2	2.2
"	Panaeolus leucophanes Berk.	1.1	2.1	3.2	4.2	1.1	2.2	1.1	1.1	.	.	+	.	2.1	2.1	.	3.2
"	- subbalteatus Berk.	.	1.1	3.2	+	3.2	2.1	.	3.2	1.1	.	2.1	2.1	.	2.1	.	2.1
"	- papilionaceus Bull.	.	.	3.2	2.2	2.2	3.2	.	2.2	5.2	.	2.1	+	2.1	.	.	3.2
"	Bolbitius titubans Bull.	+	.	2.1	3.2	.	1.1	5.2	.	1.2	+	.	1.1	.	.	2.2	1.1
"	Psilocybe coprophila Bull.	3.1	2.1	3.2	.	2.2	1.1	.	.	.	.	1.1	.	1.1	3.2	2.1	.
"	Coprinus niveus Pers.	.	.	1.1	.	+	.	+	.	1.1	.	.	2.1	2.1	5.2	.	2.1
"	- comatus Fl.Dan.	.	.	.	2.2	.	1.1	.	5.2	3.2	.	.	.	.	1.1	2.2	.
"	- narcoticus Batsch	.	.	+	+	.	.	.	.	.	.	1.1	.	.	.	2.1	.
"	Psilocybe semilanceolata Fr.	.	.	.	.	4.2	2.2	.	.	+	.	.	.	.	.	1.1	.
"	Panaeolus fimiputris Bull.	.	.	.	.	.	.	.	.	.	.	.	.	2.2	3.2	2.2	2.1
"	Coprinus rapidus Fr.	.	3.2	.	.	.	3.2	.	.	.	.	.	.	.	.	1.1	.
"	- lagopus Fr.	.	.	2.1	.	.	.	.	4.2	.	.	.	.	.	.	1.1	.
"	- subtilis Fr.	.	.	.	.	1.1	.	3.2	.	.	.	2.1	.	.	.	.	.
"	- fimetarius L.	.	.	.	.	.	2.1	2.1	.	.	.	.	.	.	+	.	.
"	Panaeolus acuminatus Fr.	.	.	.	.	.	.	.	.	.	.	.	2.1	2.1	.	2.2	.
"	Coprinus ephemerus Fr.	.	1.1	.	.	.	.	.	.	.	.	.	.	.	.	.	2.2
"	Psilocybe foenisecii Pers.	.	+	.	.	.	.	.	3.2	.	.	.	.	.	.	.	.
"	Coprinus papillatus Batsch	.	.	.	.	.	.	.	.	3.2	.	.	.	2.1	.	.	.
"	Bolbitius fragilis L.	.	.	.	.	.	.	.	.	.	3.2	.	+	.	.	.	.
Düngerzeiger:																	
Gma	Stropharia stercoraria Fr.	.	.	.	2.1	1.1	.	2.1	.	2.2	2.1	3.2	2.1	.	2.1	+	2.2
"	Psalliota bispora Lange	3.1	.	.	.	.	2.1	2.1	.	.	.	.	.	.	.	.	.
"	- villatica Brand.	.	1.1	.	.	.	.	.	.	.	.	.	.	.	.	.	.
"	Entoloma excentricum Bres.	.	.	.	.	.	.	+	.	.	.	.	.	.	.	.	.
"	Psalliota arvensis Schff.ssp. macrospora Möll. et Schäff.	.	.	.	.	.	.	.	4.1	.	.	.	.	.	.	.	.
"	Psalliota edulis Vitt.	.	.	.	.	.	.	.	.	2.1	.	.	.	.	.	.	.
"	- cretacea Fr.	.	.	.	.	.	.	.	.	1.1	.	.	.	.	.	.	.

Figure 3: Sample page from the digitized BIOfid corpus (taken from http://vl.ub.uni-frankfurt.de/biodiv/periodical/pageview/9028548).

Tafsir Dataset: A Novel Multi-Task Benchmark for Named Entity Recognition and Topic Modeling in Classical Arabic Literature

Sajawel Ahmed[1,2,3], **Rob van der Goot**[3], **Misbahur Rehman**[2], **Carl Kruse**[2], **Ömer Özsoy**[2], **Alexander Mehler**[1], **Gemma Roig**[1]
[1]Faculty for Computer Science and Mathematics, Goethe University Frankfurt
[2]Faculty for Linguistics, Cultures, and Arts, Goethe University Frankfurt
[3]Department of Computer Science, IT University of Copenhagen
{sahmed}@em.uni-frankfurt.de

Abstract

Various historical languages, which used to be lingua franca of science and arts, deserve the attention of current NLP research. In this work, we take the first data-driven steps towards this research line for Classical Arabic (CA) by addressing *named entity recognition* (NER) and *topic modeling* (TM) on the example of CA literature. We manually annotate the encyclopedic work of *Tafsir Al-Tabari* with span-based *NEs*, sentence-based *topics*, and span-based *subtopics*, thus creating the *Tafsir Dataset* with over 51,000 sentences, the first large-scale multi-task benchmark for CA. Next, we analyze our newly generated dataset, which we make open-source available, with current language models (lightweight BiLSTM, transformer-based MaChAmP) along a novel *script compression method*, thereby achieving state-of-the-art performance for our target task *CA-NER*. We also show that *CA-TM* from the perspective of historical topic models, which are central to Arabic studies, is very challenging. With this interdisciplinary work, we lay the foundations for future research on automatic analysis of CA literature.

1 Introduction

All languages deserve equal technologies. Named entity recognition (NER) and topic modeling (TM) are a crucial part of various downstream tasks in natural language processing (NLP), such as Entity Linking, Relation Extraction, and ultimately Question Answering. For such tasks, many research institutes and individual scholars put their emphasis on popular, high-resource languages like English, where there is already a large amount of previous work and resources available (Rajpurkar et al., 2018; Dzendzik et al., 2021; Cambazoglu et al., 2021). This definitely accelerates the progress of the ongoing data-driven NLP. However, many historical languages, such as Ancient Egyptian, Ancient Greek, and especially Classical Arabic (CA), which used to be the lingua franca of science and arts, have been mostly neglected by the NLP community. These languages possess large volumes of historical literature (CA: e.g. *Liber Algebrae et Almucabola*, *Canon Medicinae*, *Tafsir Al-Tabari*), which were and still are to this date relevant for many communities and societies, lay their foundations and even shape their further development. In order to perform historical analysis which are relevant for our modern age, we need to let these *forgotten* low-resource languages benefit from the wave of machine learning (ML) progress, thus making historical texts accessible to modern studies and approaching ethically an egalitarian state of NLP research.

To this end, within the project *Linked Open Tafsir* (Ahmed et al., 2022), firstly, we create the *Tafsir Dataset* by annotating the CA encyclopedic books of *Tafsir Al-Tabari* on exegetical studies of law, ethics and philosophy. This is done with respect to span-based *NEs*, sentence-based *topics* and span-based *subtopics*, thereby producing over 51,000 sentences and presenting the first multi-task benchmark for CA with three independent tasks.

Rasm + I'jam + Tashkil (Vocalized Arabic)
قَالَ اَحْمَدُ لِسَارِيةَ فِيْ مَكّةَ : كُلُوْا وَاشْرَبُوْا هَنِيْئًا
Rasm + I'jam (Standard Arabic)
قال احمد لسارية فى مكة : كلوا واشربوا هنيئا
Rasm (Skeleton Arabic)
ٯال احمد لسارٮه ڡى مكه : كلوا واسربوا هٮٮا
NER & TM Output (Skeleton Arabic)
#topic=kalam ٯال [احمد]PER ل[سارٮه]PER ڡى [مكه]LOC: كلوا واسربوا هٮٮا

Figure 1: Example for Arabic script-dependent preprocessing layers for the sentence *"Ahmed said to Saria in Mecca: eat and drink with happiness"* along NER & TM output.

Proceedings of the 29th International Conference on Computational Linguistics, pages 3753–3768
October 12–17, 2022.

Secondly, we develop a novel *script compression method* for Arabic text in order to examine its influence on the performance of neural models (see Figure 1). For this, we take the modern vocalized Arabic script and gradually transform it to its antique form of skeleton script *Rasm* from the 7th century by removing first, the vocalization marks *Tashkil* (consisting of dashes and circles), and second, the diacritic marks *I'jam* (consisting of dots), thus lowering the vocabulary size drastically by reducing the number of distinct letters from 280 (vocalized) over 28 (standard) to 16 (skeleton). From a historical critical perspective, the usage of this skeleton script is quite interesting as this was the first one to be used for documenting the text of the Quran. Thus, on a side note, by analyzing this ancient script, we shed light on the historical critical question of its readability.

Thirdly, we analyze our newly generated dataset, apply the leightweight BiLSTM (Lample et al., 2016; Ahmed and Mehler, 2018) and contrast its usage with *MaChAmp* (van der Goot et al., 2021), a toolkit for multi-task learning in NLP. This toolkit ideally fits to our multi-task benchmark, allowing us to conduct over 119 many-fold experimental setups with various Arabic pre-trained language models (LM), such as AraBERT, AraElectra, RemBERT. With these optimization steps, we produce the first major results for *CA-TM* and on top establish a state-of-the-art performance for *CA-NER* by achieving a value of up to *95.58% F1-score*.

Our work facilitates an automatic extraction of theological information so far buried in the bulk of paper manuscripts and volumes. By creating the necessary training data for tackling the task of NER and TM with various ML algorithms, we provide an open-source gold standard for the NLP community and hereby lay the foundations for future work on digitization of historical Arabic juridical and theological studies.

The remainder of the paper is organized as follows: Section 2 reviews related work, Section 3 presents the dataset, its historical source and provides details on the annotation tasks and their guidelines, Section 4 presents a sketch of the underlying methods, Section 5 reports and discusses our results, and, finally, Section 6 draw the conclusion.

2 Related Work

Not much work has been done in the field of NLP for CA as this language suffers from *resource poverty* in the ML community. For Modern Standard Arabic (MSA), there are only a handful of studies and resources open-source available. Noteworthy work specifically for MSA-NER has been done so far mainly by Benajiba et al. (2007) on *ANERCorp dataset* and by Mohit et al. (2012) on *AQMAR dataset*; both datasets along their NER models will be used as baselines here (see Table 1). Although these datasets are relatively small compared to those which are used for other languages in the community, to this date we do not have any other alternatives. For MSA-TM, again only few resources are freely available (El Kah and Zeroual, 2021), however, these are all built on modern web texts mainly from the genre of newspapers and social media. For the case of CA-TM, no prior work is known to the authors. Hence with our work, we lay the foundations for future research in this interdisciplinary field of historical NLP.

3 Tafsir Dataset: Annotation of Classical Arabic Literature

In this section, we describe the data source, the textual conversions performed to prepare the annotation task, the annotation guidelines and the annotation process itself.

3.1 Data Source: Raw Text to TEI Format

Al-Tabari *Al-Tabari*, in full *Abu Ja'far Muhammad ibn Jarir al-Tabari*, (born c. 839, Amol, Tabiristan, Iran—died 923, Baghdad, Iraq), is a religious scholar, author of enormous compendiums of early Islamic history and Quranic exegesis, who made a distinct contribution to the consolidation of Sunni thought during the 9th century. He condensed the vast wealth of exegetical and historical erudition of the preceding generations of Muslim scholars and laid the foundations for both Quranic and historical sciences. His major works were the *Exegesis of Al-Tabari* (Tafsir Al-Tabari) and the *History of Prophets and Kings*. In this study, we are focusing on his former work.

Edition of the book and TEI format Tafsir Al-Tabari has been published in various editions, the *Turki Edition* from 2001 is the most extensive and complete one, hence, this was chosen for our study. It is published in 26 volumes consisting of a total of 18,594 pages. The original text of this edition, which is vocalized, is freely available from different online sources such as the *King Saud University*, the *Shamela Software*, and from the well-known

Corpus	***Sent.***	***PER***	***LOC***	***ORG***	***TME***	***OTH***
ANERCorp-2007	5,887	3,598	*4,429*	2,231	n/a	1,115
AQMAR-2012	2,646	1,468	1,443	450	n/a	*2,474*
Tafsir-2022	**51,704**	**176,105**	**5,583**	**22,026**	**4,160**	**12,453**

Table 1: Major open-source NER Datasets for Arabic along our NER annotations in the Tafsir Dataset.

resource platform *Gawami' al-Kalim*[1], whose text is the most refined and accurate one according to a review of the linguists in our annotation team.

The raw text was transformed to the TEI format (with an adapted TEI model), which was selected due to its extensive usage in Digital Humanities (Maraoui et al., 2017). Furthermore, this format can be useful for additional data analytical inquiries (e.g. with XQuery).

Sentence splitting heuristic Sentence splitting has been addressed by various approaches (Schweter and Ahmed, 2019). However, if there is no punctuation available, it becomes challenging for many algorithms to find a stable solution. In the case of CA literature, we rarely find regular punctuation. In fact in this ancient literature, there was no concept of sentences in the modern sense. Therefore, we apply a heuristic, which first uses all possible punctuation (which are introduced by modern editing authors), then looks for some specific sense splitting words, e.g. and (*wa*), so (*fa*), then (*thumma*). With this, we achieve an average sentence length of 30 words, which proves to be useful according to our initial downstream task evaluations.

3.2 Annotation Tasks

We developed annotation guidelines for generating the Tafsir Dataset. For NER, we extended the standard task to the domain of theology. Our guidelines built on those developed for the NER dataset on German historical literature (Ahmed et al., 2019). We took the original German guideline text and adjusted it by incorporating domain-specific needs for CA. For TM, we categorized the number of topics according to the classical understanding of *tafsir studies* and its 15 fields (Al-Suyuti, 1505), and refined them further during our discussion sessions with the annotation team. The appendix shows the material which was provided to the annotation team, including the introductory example of annotations. Overall, the raw text was annotated chapter-wise by considering each verse as a single annotation task. By this scheme, we ensured that annotators had the contextual information they needed to make their interpretations.

3.2.1 Named Entities

NEs are entities that are referred to in natural language texts by proper nouns (PN) as unique individuals (e.g. Mecca, Asia, Tabari, Shia). PN are contrasted by *common names* (CN) which refer to classes of entities (e.g. city, continent, person, organization).

In our task of CA-NER, we focus on PN. However, it is not easy to differentiate between PN and CN. In the following, we provide details for each class of NE which we used to annotate our raw text (for annotation results see Table 1, for further examples of NEs see Appendix A).

Person (PER) Naming can be a complex process in classical Arab society (comparable to ancient Hebraic naming) (Almuhanna and Prunet, 2019). Full names are made of chains of single names, which can include the name of the city where the person was living. Once the full name is mentioned, short forms are usually used throughout the remainder of a text (e.g. Al-Tabari). In CA-NER, we consider all naming conventions found in the raw texts.

Location (LOC) Location names are mostly straightforward (either classical Arabic names, or names going back to ancient age of Babylonia). Sometimes, there is a ambiguity in their semantics, e.g. the word *Medina* (city) is not a PN per se, however, when it is used a short form for *Medina Al Munawwarah* ("The Enlightened City"), then it becomes a PN. Obviously, the word's meaning is highly context dependent.

Organization (ORG) We extended the modern definition of this class to the classical context of religious organizations (Jews, Christians, Muslims), their subgroups (Sunni, Shia, Ismailities), theological school of thoughts (Hanafi, Maliki, Shafi'i,

[1] https://gk.islamweb.net

Topic/Subtopic	Sent.	Span
adyan (non-Islamic relig.)	13,564	1,063
asbab (occas. of revelation)	3,086	997
fiqh (jurisprudence)	9,782	7,707
israiliyat (Judeo-Christian)	3,260	0
kalam (Islamic theology)	17,208	3,066
lugha (linguistics)	14,444	9,543
mushkilat (problem)	61	0
mutashabih (allegorical)	153	0
naskh (abrogation)	544	223
qiraat (recitation style)	1,525	2,519
sirah (prophetic biography)	1,193	215
sufism (mysticism)	7,749	881
takhsis (specification)	146	0
tikrar (repetition)	174	0
ulum (science)	2,520	823
total annotations	*75,409*	*27,037*

Table 2: Statistics for sentence-based topic and span-based subtopic annotation data.

Hanbali), tribes and clans (Hashim, Quraysh), and ethnic groups (Arabs, Greeks, Persians).

Time (TME) In the early 7th century, the moon calendar was still in its primary form, hence there was not a proper usage of numerical format like in our modern days. Therefore, dates were mostly written out in words, either only by day name, or sometimes including the month name, and rarely, the year. In CA-NER, we consider all possible variants and annotate them accordingly. Also well-known temporal entities, such as the *Day of Judgment* (Yawm Al-Din), are annotated with the tag *TME*.

Other (OTH) All NEs which did not fit into the former class were annotated with the tag *OTH*, such as name of languages (Arabic, Greek, Latin), angels (Gabriel, Michael, Raphael), and (polytheistic) deities (Al Uzza, Al Lat, Manat, Baal).

3.2.2 Topic Modeling

TM is the task of mapping (segments of) texts to a fixed set of *topics* according to a multiclass setting (Blei et al., 2003). This task is important for higher-level NLP tasks such as Semantic Search, Text Summarization and Question Answering. There is no standard number of topics, as this depends on the application domain, the desired thematic resolution and the specifics of the underlying texts. In our case of historical-exegetical tafsir studies, we determined a set of 15 sentence-based topics and span-based subtopics. Table 2 shows them along their amount of annotation data. The totals include multiple counts due to multiple annotations of the same topic. If there are lines with 0 spans and several sentences (e.g. for israiliyat), that means that only sentences have been annotated according to the 15 topics. However, no specific spans (inside the sentences) could be identified by the annotators and marked accordingly. Hence, both tasks, namely sentence-based TM and span-based TM, are displayed in Table 2, indicating that they are independent from each other.

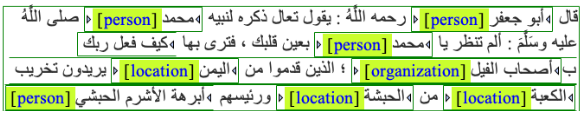
3.3 Annotation Process

Annotation Team The annotation team consisted of 4 domain experts, who were historical linguists and orientalists by background. For NER, we let the annotators train on a smaller subset of the text (i.e., chapter 50, verse 1-22) until they reached a high inter annotators agreement (IAA) value of 97% (Cohen's kappa; (Cohen, 1960)). Thus we let them continue their annotations for the remaining volumes of text individually. For TM, we did not compute any IAA value initially, as there were only 2 domain experts available for our topics. However, we ensured a high quality of topic annotation by cross-validating and correcting them directly by the other annotator.

Tool selection & issues Selecting the right tool for our annotation task was challenging. First, CA caused many problems: It is not only a low resource-language per se; even its *right-to-left* script is low-resourced to some extent, as there are not many tools that can handle it. Second, our intention was to use the TEI standard as the target data format due to its extensive usage in Digital Humanities. Third, we required a user friendly environment as our annotators did not have any technical background. Reflecting these points, we preferred the annotation tool *Oxygen XML Editor*[2] over other candidates (such as WebAnno or BRAT). Figure 2 gives a glimpse into the annotation environment.

Data format For our final training data, we use the CoNLL format (with the BIO/IOB2 tagging scheme) and extend it for the annotation of topics and subtopics. In this adjusted 3-column format, each sentence is written vertically along its *Arabic*

[2] https://oxygenxml.com/



قال أبو جعفر [person] رحمه اللَّهُ : يقول تعال ذكره لنبيه محمد [person] صلى اللَّهُ
عليه وسَلَّمَ : ألم تنظر يا محمد [person] بعين قلبك ، فترى بها كيف فعل ربك
بأصحاب الفيل [organization] ؛ الذين قدموا من اليمن [location] يريدون تخريب
الكعبة [location] من الحبشة [location] ورئيسهم أبرهة الأشرم الحبشي [person]

Figure 2: Screenshot of annotation working environment in *Oxygen XML Editor*.

token, *NE-tag* and *subtopic-tag*. Besides, for topics, a binary matrix structure is used at the beginning of each sentence to model all the occurrences of each *15 topics* (e.g. `# kalam: 1`, see sample excerpt in Appendix C).

After randomizing the order of the sentences, we divided the Tafsir Dataset into *train, dev, test* files according to the conventional ratio of 80:10:10 percentages. These resulting data files are used for our empirical evaluations, whose setups are described in the next section.

4 Methods

4.1 Script Compression

Arabic is a language with rich morphological variety of words. Besides, it has a distinct type of writing system (*Abjad*), which contains many layers of information developed in the course of the first centuries after the advent of Classical Arabic written tradition in the 7th century CE. The Arabic writing system is made of a basic skeleton script (*Rasm*), which 1-2 centuries later was extended to the standard Arabic script with the diacritic points (*I'jam*) to reduce the ambiguity of over 25 letters. Further 1-2 centuries later, the vocalization marks consisting of dashes and circles (*Tashkil*) were added which allowed a proper vocalized reading of theological literature.

Thus to deal with these variants, we propose the analytical setup shown in Figure 1. We use three textual variants, namely *skeleton*, *standard*, and *vocalized*, which denote the above mentioned stages the Arabic script went through during its historical development. We utilize the Python libraries *camel tools v1.3.1* (Obeid et al., 2020) and *rasmipy v0.2*[3], both applying rule-based preprocessing methods for generating our respective layers.

We hypothesize that $F_1(vocalized) < F_1(standard) \leq F_1(skeleton)$: The vocalization introduces noise, thus creating many different word embeddings of one word, which in turn lowers the overall *vocabulary coverage* of the LM for the training data. Hence, the standard/skeleton text will suite best to transformer-based neural models. Besides, current contextualized word embeddings are able to deal better with incoming textual data which has been the least preprocessed and overloaded with details (i.e. low feature engineering), which is the case for the standard/skeleton scripts. Moreover, for historical experts of the skeleton script, the ambiguity of each word decreases once longer contexts are provided, as they narrow down the possibilities of proper reading. Thus, we postulate that depending on the context, the model will be able to disambiguate the word itself and deliver an actual proper reading of the Arabic script. In Section 5, we will see that indeed our assumption has been right, and we find results which support this postulation.

4.2 Word Embeddings

We train word embeddings from scratch on large text corpora. For MSA, we take the *LeipzigArabic-2020* corpus (Goldhahn et al., 2012) with 13.55 Mio. sentences, which is already preprocessed such that it contains per line a sentence. For CA, we crawl the platform of *OpenITI* (Miller et al., 2018), containing the largest collection of online-available historical books for CA. Next, we apply our sentence splitting heuristic and tokenization from camel tools to produce a final text data file which again contains per line a sentence. With this, we get 134.17 Mio. sentences (with 17 GB of raw text data), the largest amount yet to be used for CA.

We calculate our optimized word embeddings with the extended version of the *Word2vec* algorithm (Mikolov et al., 2013), namely *Wang2vec* (Ling et al., 2015), with dimension 100, windows size 8, and min. word count 4. Although since 2019/2020 static word embeddings (which are context-independent after their training) are being replaced by their transformer-based generalization of pre-trained LMs, such as *BERT*, *XLNet*, *GPT-3* (which consider the context after their training), we still inspect the former method due to it allowing us to calculate a LM according to our chosen layer from Figure 1, and thus consider a *full analytical setup*. Furthermore, this allows us to examine how improvements can be achieved while using lightweight neural models, compared to data and computation intensive transformer-based LMs, which are on top expensive to train from scratch, and have a fixed vocabulary of subword units.

[3] https://pypi.org/project/rasmipy/

Data	Embeddings	skeleton	cov	standard	cov	vocalized	cov
ANERCorp	n-gram	n/a	n/a	55.23	Benajiba (2007)	n/a	n/a
AQMAR	SVM	n/a	n/a	69.33	Mohit (2012)	n/a	n/a
ANER	LeipzigAr	79.13	0.97	**79.14**	0.96	68.91	0.16
AQMAR	LeipzigAr	68.34	0.97	**70.93**	0.94	59.51	0.27
Tafsir	*OpenITI*	*87.13*	*0.99*	***87.41***	*0.99*	*82.97*	*0.52*

Table 3: BiLSTM results for NER on Tafsir Dataset for each layer (full setup). Coverage denotes the percentage of words from the training data that occur in the pre-trained embeddings.

4.3 Neural Models

This section provides details on the neural models which were used to examine the Tafsir Dataset along the script compression method.

4.3.1 BiLSTM

We use the neural model of BiLSTM-CRF (Lample et al., 2016; Ahmed and Mehler, 2018) with default hyperparameters for the task of CA-NER. In short, this model consists of stacked LSTM layers which receive the embedded tokens of an incoming sentence and compute a hidden representation, which in turn is used by the last CRF layer to predict the output NE-tags (i.e. PERson, LOCation, ORGanization, OTHers, O). For further details, we refer to the original papers.

4.3.2 MaChAmp

For our experiments with transformer-based LMs we use MaChAmp (van der Goot et al., 2021), a toolkit focused on multitask learning for NLP. We used v0.3 beta with default hyperparameters and compare all Arabic LMs we could find on the *Hugging Face* (Wolf et al., 2020) hub. In MaChAmp, each task has its own decoder, while the encoder (i.e. LM) is shared. We empirically saw that adding a CRF layer was beneficial (see Appendix E, Table 10), so we enabled it for NER as well as the subtopic task layer. Because the sentences can be annotated with multiple topics, we model each topic as a separate binary task. For the multi-task setups, we use an equal loss weight for all tasks, and process all tasks simultaneously.

5 Results

In this section, we present the results which are obtained while utilizing the methods and their setups described in the previous section. The evaluation of the NER predictions are performed by running the official evaluations script from the CoNLL 2003 shared task (Tjong Kim Sang and De Meulder, 2003) on the test set of the Tafsir Dataset.

5.1 BiLSTM Evaluation for CA-NER

In the single training setup, the results for our Tafsir Dataset is given which is preprocessed according to the layers outlined in Section 4.1. Most importantly, in contrast to transformer-based networks, this lightweight model allows us to not only process the training data according to our script compression method, but also the underlying LM of Word2vec (i.e. full setup). Table 3 shows the results for this setup.

First, we can see that the vocalized layer gives the lowest performance which confirms our original assumption. This performance is clearly linked to the low vocabulary coverage of this layer in respect to the pre-trained word embeddings on our selected corpora. Next, we see that the performance for standard and skeleton is relatively high. We can see that the skeleton layer continuously approaches the performance of the standard one. This behavior is stable across all three datasets and two languages (namely CA and MSA). This shows, that the skeleton layer is actually robust and almost as good as the standard one.

These results already demonstrate that our approach of script compression is noteworthy. Reducing the size of specific "redundant" letters does not lead to any significant reduction of the downstream performance. On the LM level, however, we save a relatively large amount of memory, e.g. for the Word2vec model calculated on the OpenITI corpus, we go down from 1.5 GB (standard) to 1.2 GB (skeleton) model size. Thus our first results on script compression appear to reveal a promising research direction.

MLM (standard)	skeleton	cov	standard	cov	vocalized	cov
aubmindlab/bert-base-arabertv02	85.37	0.87	**95.58**	1.00	80.26	0.85
aubmindlab/bert-large-arabertv2	85.13	0.86	**95.24**	1.00	80.14	0.84
CAMeL-Lab/bert-base-arabic-camelbert-ca	89.12	0.91	**95.43**	1.00	80.31	0.85
aubmindlab/araelectra-base-generator	84.94	0.87	**94.89**	1.00	80.06	0.85
bert-base-multilingual-cased^{+}	88.85	0.90	**95.15**	1.00	94.36	1.00
xlm-roberta-large^{+}	95.00	1.00	**95.29**	1.00	94.88	1.00
google/rembert^{+}	95.26	1.00	**95.32**	1.00	94.73	1.00

Table 4: MaChAmp results for NER on Tafsir Dataset with selected MLMs (all pre-trained on the standard layer), where for each layer (skeleton, standard, vocalized) its respective coverage (cov) is given.

5.2 MaCHAmp Evaluation

CA-NER In this section, we examine the Tafsir Dataset with various pre-trained Masked Language Models (MLM) from Hugging Face in over 119 multi-learning setups in MaChAmP. We start by utilizing all available Arabic MLMs (only pre-trained on the standard layer) and examining them along adding an optional CRF layer (see Appendix E, Table 10). Next, we cross test the Tafsir Dataset on the final selected MLMs, giving our major results in Table 4.

Although in respect to the script-dependent analysis, this is not the justified full setup, we can still get an idea what the impact of each script layer can be while fine-tuning the model. We see that the standard layer performs the best, confirming one part of our hypothesis that $F_1(vocalized) < F_1(standard)$ holds. Moreover, it is clearly demonstrated how the different layers influence the vocabulary coverage, which in turn influences the downstream performance. We can observe that in cases where $cov(vocalized) < cov(skeleton)$ holds, $F_1(vocalized) < F_1(skeleton)$ holds as well. In the opposite case, $vocalized$ is outperforming $skeleton$. Besides, we have noteworthy cases of MLMs marked with $^{+}$: For all these large multi-lingual models, their *word piece algorithm* is able to handle the vocalization by splitting it from each character, thus automatically producing the standard layer for the vocalized input. Last but not least, we can see that transformer-based models with an additional CRF layer outperform the lightweight BiLSTM thoroughly, even on the mismatched layers of $vocalized$ and $skeleton$. With this, we establish a state-of-the-art performance for CA-NER with 95.58% F-score. Thus, this comprehensive analysis allows researchers to use our dataset with the described model configurations to train a NER tagger that can confidently annotate related CA literature.

CA-TM & Multi-Task Learning In this setup, we fine-tuned the MLMs on the full Tafsir Dataset, first for each task separately, then joined within the setup of multi-task learning. Although the performance for CA-NER has been high, our results show that it is not beneficial for the task of CA-TM (see Appendix E, Table 11). However, multi-task learning is not always beneficial, as the cost of parameter sharing can become higher than the benefits of knowledge sharing. Besides, we hypothesize that TM is a very hard task on our unbalanced data which has many topics with small amount of training samples (see Table 2). A second reason that makes CA-TM a very challenging task is the fact that the topics were chosen mainly on the basis of normative considerations of a historical author: They should accompany the interpretation of religious texts in a normative way, so to speak, and are therefore of importance for the historical research of CA. TM has here the special task to reflect that the topics have been normatively pre-selected in a historical context that may not be directly available to contemporary annotators (for the purpose of generating appropriate training data). Nevertheless, these historical topic labels cannot simply be ignored, since they de facto shape research on CA.

5.2.1 Learning Curve over CA-NER Annotation Data

In order to evaluate the importance of our large-scale annotation work, we analyze the influence of the annotation data size on the final performance by plotting a learning curve over the annotation data. For each step of the size 5k sentences, we calculate the F1-score for CA-NER (on the test set) with the best observed model *bert-base-arabertv02*.

Figure 3 shows the learning curve displaying the downstream performance according to the progress of our annotation work.

Interestingly, we can see that the annotators' work has been worth it. The curve is quite steep, i.e. with every additional generation of annotation data we increased the performance steadily for our target task of CA-NER until 30k sentences. After that, the gradient starts to decrease at which the curve begins to slowly approach the max performance value of 95.58% F-score. Thus, we conclude that large amount of gold data is indeed beneficial for CA-NER, which contrasts previous findings for other low-resource languages such as Danish (Plank et al., 2020).

5.3 Error Analysis for CA-NER

Our manual error analysis on the test set has revealed that the following errors exist: A majority of (1) prediction errors, where the model does not tag those NEs which are annotated by the annotators, and a minority of (2) annotation errors, where the model tags those NEs which are falsely not annotated by the annotators. However, most of the annotation errors were found in the false positives.

The Arabic language contains various words with polysemy (i.e. one word has many meanings). Especially if a word is not vocalized, and the sentence context is small, it can become difficult for the common reader to understand the underlying meaning. Then, only a domain expert can provide the precise meaning. For prediction errors, our manual error analysis has shown that the model is mistaken exactly in such cases, where there is a NE in very short sentences (e.g. 2-word *nominal sentences*). We hypothesize this is because the model has only access to one sentence, whereas the domain expert annotators have more advantages by knowing the full context via their chapter-wise view.

6 Conclusion

In this work, we presented the Tafsir Dataset, the first large-scale multi-task benchmark on NER and TM in Classical Arabic literature. We demonstrated how useful resources can be for languages which have been historically important but now forgotten by the ongoing NLP research. Besides, we also performed a first evaluation of this newly generated dataset. While doing so, we empirically saw that adding a CRF layer was beneficial to

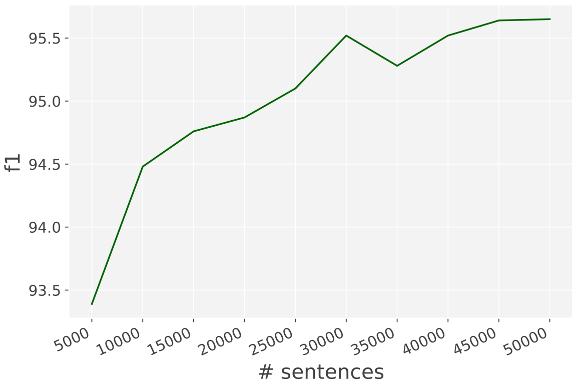


Figure 3: Learning curve over annotation data for NER (standard layer) in steps of 5k sentences.

the transformer-based models, with which we ultimately established a state-of-the-art performance for CA-NER. Although TM was not the primary focus of this paper, we generated first results for CA-TM, thereby leaving room for future improvements. This refers to a scenario of TM in which topic labels were originally determined in a historical, normative, exegetical setting, whereas they need to be learned using modern NLP tools, based on their relevance to CA research. Such scenarios are likely to be increasingly encountered as more historical languages come into NLP focus. We therefore believe that our benchmark induces a new challenge for the NLP community that can lead to progress for our target low-resource language.

The Tafsir Dataset and its accompanying material are made open-source available for the research community. Furthermore, a website[4] is published which offers a comprehensive research tool in English and Arabic for accessing our dataset in a more user-friendly environment and performing various search queries on it. The web-based tool is freely available and provides over 400 filter options along the categories of our dataset. Additionally, it provides the option of graphical visualization (bubble or pie chart) of the dataset and of the query results performed on it. This digital tool makes it possible for scholars from historical and theological fields to access the dataset without any prior technical skill sets, thus allowing them to find systematically the answers to their long-lasting research questions.

On a side note, by analyzing the historical skeleton script, we shed light on a centuries-old historical critical question regarding the readability of the Rasm text: Whether the first Quranic manuscripts

[4] https://linkedopentafsir.de/

(i.e. *Uthmanic codex*) can provide a precise reading of the canonized oral text, or whether there is a large amount of ambiguity in it. Our script-dependent analysis shows that from an information retrieval perspective, the usage of the skeleton script is robust enough to deliver a similar performance compared to the usage of the standard script. We can thus conclude that if the ML model is able to deal with the skeleton script, then humans will also not face major difficulties after gaining sufficient training on the same ancient script.

Future work Our work gives indications that script compression seems to be a promising direction to reduce the amount of data and tackle the question of which resource-size actually matters (Ahmed and Mehler, 2018). In this work, for the case of Arabic we came down from 28 to 16 letters while keeping the performance stable. This shows that we do not need (1) vowels, and (2) different letters for each phoneme. In fact, just some minimum amount of *consonantal distinction* is needed. What is this amount, can we determine it exactly for each target language? Phonetic algorithms such as *Metaphone* (Philips, 1990) pose to be a first language-independent approach, be that as it may, only future work can give us the answers.

Acknowledgments

This interdisciplinary research work has been conducted as part of the project *Linked Open Tafsir*, which is funded by the *AIWG*[5], in turn supported by the *BMBF* and *Stiftung Mercator*. We acknowledge the IT University of Copenhagen HPC for the resources made available for conducting the research reported in this paper. Special thanks go to U. Brucker and the International Office (Goethe University Frankfurt) for enabling the collaboration with the IT University of Copenhagen, Prof. B. Plank and further colleagues from the IT University of Copenhagen for their critical feedback on the experimental evaluations, Prof. U. Meyer and GRADE (Goethe University Frankfurt) for funding the final publication process, and Prof. G. Hirst (University of Toronto) along Dr. J. Kindermann (Fraunhofer Institute IAIS) and Prof. R. V. Zicari (Goethe University Frankfurt) for their closing remarks on the manuscript.

[5] https://aiwg.de/

References

Sajawel Ahmed and Alexander Mehler. 2018. Resource-Size matters: Improving Neural Named Entity Recognition with Optimized Large Corpora. In *Proceedings of the 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Florida, Orlando, USA.

Sajawel Ahmed, Misbahur Rehman, Joshua Tischlik, Carl Kruse, Edin Mahmutovic, and Ömer Özsoy. 2022. Linked Open Tafsir—Rekonstruktion der Entstehungsdynamik (en) des Korans mithilfe der Netzwerkmodellierung früher islamischer Überlieferungen. In *8. Jahrestagung des Verbandes Digital Humanities im deutschsprachigen Raum (DHd)*.

Sajawel Ahmed, Manuel Stoeckel, Christine Driller, Adrian Pachzelt, and Alexander Mehler. 2019. BIOfid dataset: Publishing a German gold standard for named entity recognition in historical biodiversity literature. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 871–880, Hong Kong, China. Association for Computational Linguistics.

Jalal Al-Din Al-Suyuti. 1505. *Al-itqan Fi 'ulum Al-Qur'an (The Perfect Guide to the Sciences of the Qu'ran)*. Garnet Publishing; Bilingual edition (May 1, 2012).

Amin Almuhanna and Jean-Francois Prunet. 2019. From Classical to Modern Arab Names and Back. *Anthropological Linguistics*, 61(4):405–458. Copyright - Copyright University of Nebraska Press Winter 2019; Last updated - 2021-10-04; SubjectsTermNotLitGenreText - Arabian Peninsula; Kuwait.

Yassine Benajiba, Paolo Rosso, and José Miguel Benedíruiz. 2007. Anersys: An arabic named entity recognition system based on maximum entropy. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 143–153. Springer.

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Armin Burkhardt. 2004. 2004. Nomen est omen? : zur Semantik der Eigennamen. In *Landesheimatbund Sachsen-Anhalt e. V. (Hrsg.): "Magdeburger Namenlandschaft" : Orts- und Personennamen der Stadt und Region Magdeburg*.

B Barla Cambazoglu, Mark Sanderson, Falk Scholer, and Bruce Croft. 2021. A review of public datasets in question answering research. In *ACM SIGIR Forum*, volume 54, pages 1–23. ACM New York, NY, USA.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Daria Dzendzik, Jennifer Foster, and Carl Vogel. 2021. English machine reading comprehension datasets: A survey. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8784–8804, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Anoual El Kah and Imad Zeroual. 2021. Arabic topic identification: A decade scoping review. In *E3S Web of Conferences*, volume 297, page 01058. EDP Sciences.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *LREC*.

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. Association for Computational Linguistics.

Wang Ling, Chris Dyer, Alan Black, and Isabel Trancoso. 2015. Two/Too Simple Adaptations of word2vec for Syntax Problems. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Hajer Maraoui, Kais Haddar, and Laurent Romary. 2017. Encoding prototype of al-hadith al-shareef in tei. In *International Conference on Arabic Language Processing*, pages 217–229. Springer.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Matthew Thomas Miller, Maxim G. Romanov, and Sarah Bowen Savant. 2018. Digitizing the textual heritage of the premodern islamicate world: Principles and plans. *International Journal of Middle East Studies*, 50(1):103–109.

Behrang Mohit, Nathan Schneider, Rishav Bhowmick, Kemal Oflazer, and Noah A Smith. 2012. Recall-oriented learning of named entities in arabic wikipedia. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 162–173.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. CAMeL tools: An open source python toolkit for Arabic natural language processing. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 7022–7032, Marseille, France. European Language Resources Association.

Lawrence Philips. 1990. Hanging on the metaphone. In *Computer Language*, volume 7, pages 39–43.

Barbara Plank, Kristian Nørgaard Jensen, and Rob van der Goot. 2020. DaN+: Danish nested named entities and lexical normalization. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6649–6662, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Stefan Schweter and Sajawel Ahmed. 2019. Deep-EOS: General-Purpose Neural Networks for Sentence Boundary Detection. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS)*.

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 the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021. Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Examples for annotating named entities in Tafsir Al-Tabari books

1) ذُكر أن [المشركين]ORG سألوا رسول الله صلى الله عليه وسلم عن نسب ربّ العزّة, فأنـزل الله هذه السورة جوابا لهم. (112:1)
2) حدثنا [أبو كُرَيب]PER , قال: ثنا [ابن إدريس]PER , عن [عبد الملك]PER , عن [طلحة]PER , عن [مجاهد]PER , مثله. (112:4)
3) يقول تعالى ذكره لنبيه [محمد]PER صلى الله عليه وسلم : إذا جاءك نصر الله يا [محمد]PER على قومك من [قريش]ORG , والفتح: فتح [مكة]LOC (110:2)
4) قال: ثنا [ابن ثور]PER ، عن [معمر]PER ، عن [ابن طاوس]PER ، عن أبيه، قال: ما من شيء أقرب إلى الشرك من رُقْية المجانين . (113:4)
5) حدثني [يعقوب]PER ، قال: ثنا [هشيم]PER ، قال: أخبرنا [العوّام بن عبد الجبار الجولاني]PER ، قال: قَدم رجل من [أصحاب رسول الله]ORG صلى الله عليه وسلم [الشأم]LOC ، قال: فنظر إلى دور [أهل الذمة]ORG ، وما هم فيه من العيش والنضارة، وما وُسّع عليهم في دنياهم، قال: فقال: لا أبا لك أليس من ورائهم الفلق؟ قال: قيل: وما الفلق؟ قال: بيت في [جهنم]LOC إذ فُتح هَرّ [أهْلُ النار]ORG . (113:1)
6) يقول تعالى ذكره لنبيه [محمد]PER صلى الله عليه وسلم: ألم تنظر يا [محمد]PER بعين قلبك, فترى بها كَيْفَ فَعَلَ رَبُّكَ بِ [أَصْحَابِ الْفِيلِ]ORG الذين قَدِموا من [اليمن]LOC يريدون تخريب [الكعبة]LOC من [الحبشة]LOC ورئيسهم [أبرهة الحبشيّ الأشرم]PER (105:1)
7) حدثنا [ابن المثنى]PER ، قال: ثني [عبد الأعلى]PER ، قال: ثنا [داود]PER ، عن [عكرِمة]PER ، عن [ابن عباس]PER ، قال: نـزل [القرآن]OTH كله مرة واحدة في [ليلة القدر]TME في [رمضان]TME إلى السماء الدنيا، فكان الله إذا أراد أن يحدث في [الأرض]LOC شيئًا أنـزله منه حتى جمعه. (97:1)

Figure 4: Examples for annotating named entities (i.e. PER, LOC, ORG, TME, OTH) in 7 verses from the raw text of Tafsir Al-Tabari books.

B Annotation Guidelines (German version)

Guidelines für die Named Entity Recognition. Sie bauen auf den arabisierten Guidelines von Ahmed et al. (2019) auf.

B.1 Einführung: Named Entity Recognition

Unter der Named Entity Recognition (NER) versteht man die Aufgabe, Eigennamen (named entities) in Texten zu erkennen. Technisch gesehen sind hierzu zwei Schritte notwendig. Zuerst müssen in einem laufenden Text die Token gefunden werden, die zu einem Eigennamen gehören (Named Entity Detection: NED), danach können diese Eigennamen semantischen Kategorien zugeordnet werden (Named Entity Classification). Prototypisch ist dabei der Unterschied zwischen Eigennamen und Appellativa der, dass letztere eine Gattung oder eine Klasse beschreiben, während erstere einzelne Individuen oder Sammlungen von Individuen unabhängig von gemeinsamen Eigenschaften bezeichnen (Burkhardt, 2004). Die vorliegenden Guidelines sollen es Annotatoren ermöglichen, Eigennamen in Texte aus Standard und Nichtstandard-Varietäten konsistent zu annotieren. In diesen Guidelines werden die beiden Aufgaben der NED und NEC nicht unterschieden, da die Konzentration auf Beispiele in diesem Dokument, die Trennung künstlich erzeugen müsste und nicht zu erwarten ist, dass die Resultate sich dadurch verbessern würden. In Anlehnung an die oben genannten Guidelines für Zeitungssprache werden in NoSta-D-Tafsir fünf semantische Hauptklassen für klassiche arabische Texte unterschieden (Personen, Organisationen, Orte, Zeiten und Andere).

B.2 Wie finde ich eine NE?

Schritt 1: Nur volle Nominalphrasen können NEs sein. Pronomen und alle anderen Phrasen können ignoriert werden.
Schritt 2: Namen sind im Prinzip Bezeichnungen für einzigartige Einheiten, die nicht über gemeinsame Eigenschaften beschrieben werden.
Beispiel:
[Der Struppi] folgt [seinem Herrchen].
Hier gibt es zwei Nominalphrasen als Kandidaten für einen Eigennamen (NE). “Der Struppi” bezeichnet eine einzige Einheit. Es kann auch mehrere Struppis geben, aber diese haben an sich keine gemeinsamen Eigenschaften, bis auf den gemeinsamen Namen, daher handelt es sich um einen Eigennamen. *"seinem Herrchen"* bezeichnet zwar (typischerweise) auch nur eine einzige Person allerdings können wir diese nur über die Eigenschaft identifizieren, dass sie ein Herrchen ist und dass dies für Struppi zutrifft. Struppi könnte auch mehrere Herrchen haben, die alle die Eigenschaften teilen, die ein Struppi-Herrchen beinhaltet (z.B. darf Struppi streicheln, muss ihn ausführen und füttern etc.)
Schritt 3: Determinierer sind keine Teile des Namens.
Beispiel: *Der [Struppi]NE folgt seinem Herrchen.*
Schritt 4: Eigennamen können mehr als ein Token beinhalten. Beispiel:
Viele Personennamen (PER für person):
[Abu Jafar Muhammad Ibn Jarir Al Tabari]PER
Buchtitle (OTH für other):
[Jami Al Bayan Al Tawil Ay Al Quran]OTH
Schritt 5: Eigennamen können auch in einander verschachtelt sein. Beispiel:

Personennamen in Buchtiteln:
[Sunan [Abi Dawud]`PER`*]*`OTH`
Orte (LOC für location) in Vereinsnamen (ORG für organisation):
[Hebarium Senckenbergianum [Frankfurt]`LOC`*]*`ORG`
Schritt 6: Titel, Anreden und Besitzer gehören NICHT zu einem komplexen Eigennamen. Besitzer können natürlich selber Eigennamen sein.
Beispiel:
Referenz auf Musiktitel:
[Vivaldis]`PER` *[Vier Jahreszeiten]*`OTH`
Referenz auf Personen:
Landesvorsitzende Frau Vorstandsvorsitzende Dr. [Ute Wedemeier]`PER`
Schritt 7: Wenn das Gesamttoken einen Eigennamen darstellt, dann wird dieser annotiert. Beispiel:
Stiftungen: *[[Böll]*`PER`*-Stiftung]*`ORG`
Schritt 8: Kann in einem Kontext nicht entschieden werden, ob eine NP sich als Eigennamen oder Appellativ verhält, wird es nicht als NE markiert. Beispiel:
Ortsnamen vs. -beschreibungen:
...und zogen mit ihren grossen Transparenten gestern vom [Steintor] über den [Ostertorsteinweg]`LOC` *zum [Marktplatz].*
Schritt 9: Wenn ein Name als Bezeichnung für bestimmte Gegenstände in die Sprache übergegangen ist und in seiner Nutzung nicht als NE fungiert, so wird dieser nicht annotiert. Beispiel:
[Teddybär] *(*`NICHT PER`*)*
[Colt] *(*`NICHT PER`*)*
Schritt 10: Bei Aufzählungen mit Hilfe von Bindestrichen oder Vertragen eines Teils der NE auf spätere Wörter, wird die NE so annotiert, als sei sie voll ausgeschrieben.
Beispiel:
[Frühe]`OTH` *und [Späte Bronzezeit]*`OTH`
[Süd-]`LOC` *und [Nordafrika]*`LOC`

B.3 Zu welcher semantischen Klasse gehört ein Eigenname?

Wenn der Eigenname in eine der Klassen in der Liste Faustregel zur Unterscheidung einer Klassenbezeichnung und eines Namens gehört, dann annotiere die zugehörige Klasse. Sollte die gefundene NE Rechtschreibfehler enthalten, wird sie dennoch annotiert. In Zweifelsfällen hilft auch die Tabelle NoSta-D-Tafsir-TagSet und alle Untertabellen, insbesondere die Beispiele mit dem weiter.
Jahreszahlen in ORGanisationen werden markiert. Beispiel:
[COLING]`ORG` *[2022]*`TIME`
[Fussball-WM]`ORG` *[2014]*`TIME`
Wenn der Eigennamen in KEINE der vorhandenen Klassen passt, markiere diesen mit ***UNCLEAR***, notiere dir bitte das Beispiel und schicke uns eine E-Mail an: `X.Y@email.com`. So können wir die Guidelines sukzessiv verbessern.

B.4 Faustregel zur Unterscheidung einer Klassenbezeichnung und eines Namens:

- Elemente der fraglichen Einheit verbinden die gleichen Eigenschaften → Klasse → keine NE
- Die Elemente der fraglichen Einheit verbindet nur der Name oder Element ist Einheit bezeichnet ein spezifisches Individuum → Name → NE
- *"Paleocene"* bezeichnet spezifische Epoche → NE (TME)

NoSta-D-Tafsir-Tagset

Table 5: Kategorie 'PER-Person'

Subkategorie	**Beispiele**
Person	*Ibn Ahmed, Saria, Al Tabari*
Künstlernamen	*Abu Nuwas*
Charaktere	*Ali Baba*
Superhelden	*Aladin, Sindbad*

Table 6: Kategorie 'LOC-Ort'

Subkategorie	**Beispiele**
Bezirke	*Makkah Aziziyah, Schöneberg*
Sehenswürdigkeiten, Moscheen	*Mada'in Saleh, Al Masjid Al Haram*
Planeten	*Erde, Mars*
Landschaftsbezeichnungen	*Al Nefud, Königsheide*
Straßen, Plätze	*Al Tariq Al Maliki Al Farsi*
Einkaufszentren	*Suq Ukadh, Nordwestzentrum*
Berge, Seen, Flüsse	*Jabal Arafat, Al Bahr Al Ahmar, Wadi Hanifa*
Kontinente	*Asien, Europa*
Länder, Staaten	*Saudi-Arabien, Hessen, Iran*
Städte	*Mekka, Babylon*
Regionen	*Al Hijaz*
Qiraat-Orte	*Al Amsar*

Table 7: Kategorie 'ORG-Organisation'

Subkategorie	**Beispiele**
Organisationen	*Ahl Al Hadith, Sunni, Shia, Ismailiten, GEFIS, EU, Landgericht Frankfurt*
Religionsgruppen	*Juden, Christen, Muslime*
Unternehmen	*Karimis, Microsoft*
Sammelbezeichung	*Umran*
Madhahib	*Kufiyun*
Qabilah	*Quraish*
Volksgruppen	*Araber, Perser, Römer*
Universitäten	*Al-Azhar University*
Bibliotheken	*Bayt Al Hikmah*

Table 8: Kategorie 'TIME'

Subkategorie	**Beispiele**
Tag	*Freitag*
Monat	*Rabi' Al Awwal*
Jahr	*570*
dd.mm.yyyy	*12.03.0570*
Jahrhundert	*5. Jahrhundert*
Epochen	*Jahiliyyah, Paleocene*

Table 9: Kategorie 'OTH-Andere'

Subkategorie	**Beispiele**
Buch-, Filmtitel etc.	*Sahih Al Bukhari, Faust*
Währungen	*Dinar, Dirham, Euro*
Sprachen	*Arabisch, Deutsch, Latein*
Buchtitel mittels Autor	*Helbig et al., ([[Helbig]PER et al.]OTH)*
Gottheiten	*Al Uzza, Al Lat, Manat, Ba'al, Nasr, Suwa', Wadd, Yaghuth*
Engel	*Jibril, Mikail, Israfil*
Dschinn	*Iblis*
Mythol. Tiere	*Hudhud*

C Sample Excerpt from Tafsir Dataset

```
# adyan: 0
# asbab: 0
# fiqh: 0
# kalam: 1
# lugha: 1
# mushkilat: 0
# mutashabih: 0
# naskh: 0
# qiraat: 0
# science: 0
# sirah: 0
# sufism: 0
# takhsis: 0
# tikrar: 0
وَقَوْلُهُ     O     O
:          O     O
وُجُوهٌ      O     O
يَوْمَئِذٍ     O     O
نَاضِرَةٌ     O     O
.          O     O
يَقُولُ      O     O
تَعَالَى     O     O
ذكره       O     O
:          O     O
وُجُوهٌ      O     B-lugha
يَوْمَئِذٍ     O     O
.          O     O
يَعْنِي      O     O
:          O     O
يَوْمَ       B-TME O
الْقِيَامَةِ   I-TME O
،          O     O
نَاضِرَةٌ     O     B-lugha
.          O     O
```

Figure 5: Tafsir Dataset in CoNLL format, showing the binary topic matrix before the sentence start, afterwards the Arabic tokens along their NER tag (1st column) and subtopic tag (2st column).

D Data Statement

In accordance with (Bender and Friedman, 2018), the following outlines the data statement for the Tafsir Dataset:

A. CURATION RATIONALE Manual annotation of literature in Classical Arabic, which is to date a low-resource language, for identification of named entities in different historical text domains, complemented with topic modeling annotation. The generation of such training data enables machine learning applications for the research fields of historical NLP and digital humanities.

B. LANGUAGE VARIETY The canonical text data of *Tafsir Al-Tabari* was collected from the online resource platform *Gawami' al-Kalim* (https://gk.islamweb.net).

C. SPEAKER DEMOGRAPHIC For various text samples in the historical collections of narrations, it is Classical Arabic speakers. Gender, age, race-ethnicity, socioeconomic status can be inferred from the extensive classical literature of biographical evaluation (*'Ilm Al-Rijal*) on narrators and their biographies (books such as *Al-Tarikh Al-Kabir* ("The Great History") by Imam Bukhari, *Kitab Al-Tabaqat Al-Kabir* ("The Book of the Major Classes") by Ibn Sa'd, or *Ikhtiyar Ma Rifat Al-Rijal* ("The Selection of the Knowledge of the Men") by Shaykh Tusi).

D. ANNOTATOR DEMOGRAPHIC Four scientific staff members and two students (age range: 25-60), gender: male and female. European with Middle Eastern background. Native language: German, Modern Standard Arabic, Classical Arabic. Socioeconomic status: university faculty and higher-education student in Classical Arabic studies.

E. SPEECH SITUATION Sopken Classical Arabic, which was later edited by the collector (here: Al-Tabari). Time frame of data between 7th century and 923 CE. Place: Middle East.

F. TEXT CHARACTERISTICS Exegetical literature: Sentences made of chain of narrators (*Isnad*) and the actual content of narrations (*Matn*) along exegetical prose elaborations for each verse of the Quran.

PROVENANCE APPENDIX N/A

E Extended Results

MLM	**SEQ**	**CRF**	**Coverage**
aubmindlab/bert-base-arabert	79.34	79.91	0.74
aubmindlab/bert-base-arabertv01	79.49	80.07	0.65
aubmindlab/bert-base-arabertv02	79.81	80.26	0.85
aubmindlab/bert-base-arabertv2	79.43	80.14	0.84
aubmindlab/bert-large-arabertv2	79.18	80.29	0.84
asafaya/bert-base-arabic	94.99	95.31	1.00
asafaya/bert-mini-arabic	94.02	94.50	1.00
asafaya/bert-large-arabic	94.90	94.92	1.00
asafaya/bert-medium-arabic	94.93	94.87	1.00
CAMeL-Lab/bert-base-arabic-camelbert-ca	79.56	80.31	0.85
CAMeL-Lab/bert-base-arabic-camelbert-mix	79.61	80.19	0.85
CAMeL-Lab/bert-base-arabic-camelbert-msa	79.40	80.23	0.85
UBC-NLP/ARBERT	95.04	95.29	0.88
UBC-NLP/MARBERT	94.83	94.92	0.88
aubmindlab/araelectra-base-generator	79.37	80.06	0.85
bert-base-multilingual-cased	93.89	94.36	1.00
xlm-roberta-base	94.13	94.49	1.00
xlm-roberta-large	94.36	94.88	1.00
google/rembert	94.43	94.73	1.00

Table 10: Results for CA-NER w/ and w/o CRF

MLM	**NER**		**Topic**		**Subtopic**	
	st	mt	st	mt	st	mt
aubmindlab/bert-base-arabertv02	95.99	95.87	26.11	13.73	21.18	20.47
aubmindlab/bert-large-arabertv2	95.53	95.26	18.94	14.43	18.28	19.44
asafaya/bert-base-arabic	95.61	94.94	20.63	11.84	19.23	18.36
asafaya/bert-large-arabic	95.65	95.80	22.15	20.46	21.68	20.58
asafaya/bert-medium-arabic	95.13	95.17	20.15	9.46	18.67	17.45
CAMeL-Lab/bert-base-arabic-camelbert-ca	96.06	95.99	24.75	15.81	19.68	17.42
UBC-NLP/ARBERT	95.46	95.45	22.16	20.37	22.05	20.56
aubmindlab/araelectra-base-generator	95.08	94.95	18.92	6.86	14.85	14.65
bert-base-multilingual-cased	95.04	94.79	23.11	11.58	18.54	16.72
xlm-roberta-large	95.54	95.22	16.97	13.80	21.18	20.46

Table 11: Multi-task learning results for each task. st=single task, mt=multitask

Topic	**Macro-F1**
adyan (non-Islamic religion)	27.93
asbab (occasions of revelation)	22.74
fiqh (jurisprudence)	16.66
israliyat (Judeo-Christian)	23.17
kalam (Islamic theology)	26.61
lugha (linguistics)	30.06
mushkilat (problem)	19.97
mutashabih (allegorical)	20.00
naskh (abrogation)	19.76
qiraat (recitation style)	*41.45*
sirah (biography)	21.96
sufism (mysticism)	14.87
takhsis (specification)	19.99
tikrar (repetition)	19.98
ulum (science)	18.41

Table 12: Fine-grained TM results obtained with the measure of Macro-F1 from MaChAmp on Tafsir Dataset (arabertv02).

NE category	**Precision**	**Recall**	**F1**
PER	*97.12*	*97.60*	*97.36*
LOC	72.53	66.93	69.62
ORG	82.00	89.31	85.50
TME	78.00	79.90	78.94
OTH	79.59	76.38	77.95

Table 13: Fine-grained NER results obtained by running the official CoNLL-2003 script on Tafsir Dataset (arabertv02).

4

Transformation Procedure for Data Driven Enrichment of Low-Resource Languages

In the main part of our dissertation, we have seen a step by step demonstration of how low-resource languages can benefit from the process of resource generation and optimization. We have seen that gathering unlabeled text is quite important (cf. Module 1, Figure 1.1). In particular, we have seen that the annotation work (cf. Module 2, Figure 1.1) plays a vital role in closing the gap between modern high-resource and historical, low-resource languages (cf. Module 3, Figure 1.1).

In the following, by summarizing the approach of our study papers, we present a generalized *transformation procedure* for data driven enrichment of low-resource languages (see Procedure 1). This procedure consists of six major steps and can be used as a guideline by researchers with similar research goals.

Procedure 1 transforming a given language L from low to high-resource state

```
 1: GATHER all available unlabeled text resources to form large text corpus
    $corpusData_L$ for unsupervised LM training
 2: GATHER all available labeled training data $goldData_L$ for the target NLP task
    (e.g. NER) and its supervised training
 3: DEVELOP an intermediate language-specific preprocessing module $proc_L$
 4: while true do
 5:     TRAIN $model_E(proc_L(corpusData_L), proc_L(goldData_L))$
 6:     if $F_1Score(L) \geq F_1Score(English)$ then
 7:         EXIT
 8:     end if
 9:     GENERATE annotation data $d_L^*$
10:     $goldData_L += d_L^*$
11: end while
```

Step 1: In the first step, the aim is to build a foundational text corpus $corpusData_L$ for the given low-resource language L which enables a training of the underlying language model (LM) from scratch. This is done by gathering all unlabeled digital pieces of texts which are open-source available. In some cases, these texts can only be found in large scattered collections of (OCR-generated) PDFs, thus a removal of meta-data and extraction of sentences becomes the first task while preparing the final corpus. In our study papers, we were fortunate to have access to previous work with the *COW corpus* [Sch15] for the German analysis and with *OpenITI corpus* [MRS18] for the Arabic counterpart, which had accelerated our research work strongly.

Step 2: It is possible to find the required labeled data $goldData_L$ sometimes hidden in larger treebanks or complex databases, that have not come under the focus of NLP researchers, as it has been the case for *Tübingen Treebank* for German NER (cf. *Paper 1* [ASM18]). It should be noted, however, that usually it is quite difficult to find any (open-source) dataset for a given low-resource languages (as per definition they lack digital resources). Hence the size $goldData_L$ is often quite small.

Step 3: Apart from multilingual models, most neural network models for supervised learning tasks are made in regard to high-resource languages such as English (an analytic language with low level of inflections [McA92]). However, many low-resource languages, especially the historical ones, have different grammatical and stylistic properties (e.g. heavy usage of inflections, lack of sentence punctuation, role of written text versus *oral traditions* [Ahm+22]), and do not fit directly to modern English-tailored models. Consequently, it is important to develop an intermediate language-specific preprocessing module (by remembering the paradigm *simpler is better*) before proceeding with any neural network training (cf. *Paper 1* [ASM18], *Paper 4* [Ahm+22]).

Step 5: In this step, the actual neural training is performed. As explained in the previous step, we assume that we use a neural network model $model_E$ initially made for the English language and perform the LM training as well as the task-specific training on $corpusData_L$ and $goldData_L$, respectively. Ideally, both parts are trained from scratch, however, for data-intensive models (such as BERT), the LM training usually cannot be conducted from scratch by researchers with limited compute resources. Consequently, they have to search for a pre-trained model as an alternative which is the closest match to the target language L (as it was the case in our Classical Arabic analysis, where we found pre-trained models on Modern Standard Arabic on HuggingFace [Ahm+22]).

Step 6: This decisive step measures the actual performance of the given NLP tasks (e.g. NER) and compares it with the state-of-the-art performance on the same task for a high-resource language. For simplicity, we choose English as the high-resource language.

Step 9: This is the main iterative step of our transformation procedure. We generate manually new annotation data d^*_L and add it to our existing dataset $goldData_L$ until we approach the performance of a high-resource language. Once this is the case, we have practically closed the gap to high-resource languages and thus can stop the annotation work.

Overall, we can see that the crucial part of this transformation procedure is the generation of domain-specific annotation data. This in turn requires writing precise *annotation guidelines* for human annotators, such that they can produce annotation data of high quality. Only then it is possible to achieve a high performance for a given NLP task and close the gap to high-resource languages. This was demonstrated in the German part of our dissertation, and reaffirmed in the Arabic counterpart. We thus believe, that our procedure can be applied to any human language from any domain and age.

5 Conclusion

5.1 Answers to Research Questions

In our current time, many researchers from the field of ML-based NLP focus on deep and wide models of neural networks. The more the research continues, the deeper and wider the networks become. Within such a development, low-resource languages are forgotten, leading to major scientific as well as ethical issues. In our dissertation, we have shown that by focusing on the data side of the NLP progress, this development can be countered: historical, low-resource languages can benefit from the current NLP methods the same way modern, high-resource languages like English and Chinese are benefiting from it. In this way, an automatic analysis of historical literature can be accomplished. We choose the way of data-driven enrichment of low-resource languages for a given NLP task without any network designing. In Table 5.1, we provide the final answers to our key research questions (cf. Chapter 1.2), according to the main findings of our study papers.

RQ	Answers
1	Through the example of NER, we have shown that the performance gap can only be closed by careful resource generation & optimization, i.e. by producing domain-specific high-quality (un-) labeled data for enabling a training of ML models from scratch.
2	It is not necessary to develop a new network design for each single (low-resource) language. Instead, an adjustment of the preprocessing (or the first layer of the English models) has to be conducted. Thus, our work indicates that historical languages have the same inherent linguistic structures as modern languages.
3	The more (un-) labeled text data we have, the better downstream-task evaluations become. In that way, resource-size indeed matters!
4	The fewer diacritic and vocal markers are used, the better downstream-task evaluations become. In that way, resource-quality matters more!

Tab. 5.1.: Answers to our key research questions (RQ) according to the main findings of this paper-based dissertation.

5.1.1 Discussion

This doctoral dissertation contributed to various fields of science. Our work demonstrated whether current ML methods are appropriate for processing large collections of historical texts on both biodiversity and theology. For these respective fields, it initiated the development of freely available datasets together with online search tools that allow domain researchers without any technical background to conduct their long-lasting research queries. This has been the case for historical researchers on biodiversity with the online platform *BIOfid*[1], as well as for historical researcher on religious studies with the online platform *Linked Open Tafsir*[2]. Thus, it enabled the access to valuable, so far untouched textual data in such a way that can possibly lead to new breakthroughs in the research both on climate change and on theological literary studies.

Furthermore, our work extended the ML methods to the specific needs which arise in the contexts of important, but, still under-resourced languages such as German and Arabic. We have seen that, despite the cultural, historical, and linguistic differences of our two target low-resource languages from the Indo-European and Afro-Asiatic language families, a common transformation procedure leads to comparable a improvement. With this research work, we have made steps towards countering the English-dominated landscape of NLP. There are definitely many advantages to using the English language as the primary medium of research. It is the lingua franca of the modern age, therefore allowing people worldwide to conduct their research with this language. In addition, it is an analytic language within the linguistic topology, therefore it exhibits low level of inflections [McA92], which certainly makes it better suitable to current ML methods than synthetic languages. In spite of that, we might cross new frontiers of NLP while working with different, more complex languages (like Latin, Classical Arabic, Sanskrit, or even Esperanto) that can provide new angles to old research questions. All in all, our work contributed to the advancement of artificial intelligence to human-like performance, and thereby reached new frontiers in the emerging field of historical NLP.

5.2 Future Work

5.2.1 Script-Compression

Our dissertation has shown that the method of script-compression, which allows a reduction of text data size while maintaining the performance stability, appears to be

[1] https://www.biofid.de/

[2] https://www.tafsirtabari.com/

a promising research direction for future work. We have shown that this method is practically useful for the case of Arabic language along its temporal variants, namely both for Modern Standard Arabic and Classical Arabic [Ahm+22]. In that case, we reduced the number of distinct letters from 28 to 16, thereby reducing the length of the majority of words.

Along the same research line, we believe that script-compression can be applied to Hebrew as well, as the related writing system of this Afro-Asiatic language has a similar structure. The Hebrew script appears to have undergone analogical historical developments cognate with Arabic, thus its "skeleton" script (derived from the Imperial Aramaic script) was adjusted later by scribes according to language-dependent needs by adding the layer of diacritic markers *Dagesh* (i.e. phoneme modification, similar to *I'jam*), and later by another layer of vowel markers, *Niqqud* (i.e. vowel modification, similar to *Tashkil*)[3]. Given that analogy, we postulate that future work can show that script-compression can lead to similar benefits for Hebrew NLP.

Furthermore, our work has shown that we do not need (1) specific letters for vowels, and (2) specific letters for each phoneme. In fact, only a minimum amount of *consonantal distinction* is needed. Now a generalized question can be posed for future work: What is that minimum numbers of letters, can it be determined for a given (low-resource) language? For Arabic we have determined that this number is actually around 16, thus the original skeleton script seems to be sufficient for that language from an ML perspective. For other languages, especially those which use Latin-based alphabets and therefore make extensive use of vowel letters, phonetic algorithms such as *Metaphone* [Phi90] could be a first language-independent approach. With this, we can possibly reduce the size of various words of high-resource languages such as English, thus reducing the text data size while retaining the performance stability. Be that as it may, only future work will give us the precise answers to these research questions.

5.2.2 Automatization of Transformation Procedure

For future work on low-resource languages, the question arises of whether it is possible to automatize our transformation procedure described in Chapter 4. We make the following assumption: Yes, it is indeed possible to automatize the transformation procedure with the help of current state-of-the-art machine translation (MT) tools, such as *Google Translate*[4] [Wu+16; Bap+22], *DeepL*[5], or *Bing Microsoft*

[3] Example for the word Aliza: אליזה (*skeleton*) ⇒ אליזה (*standard*) ⇒ אֲלִיזָה *(vocalized)*

[4] `https://translate.google.com/`

[5] `https://www.deepl.com/translator`

Translator[6]. Given that their translation performance is strong enough (Bleu Score $>=$ 0.7) [Pap+02] for the target historical language, with such MT tools, both the unlabeled and especially the labeled text data can be translated, avoiding any time and cost-intensive annotation work.

However, we should know the limitations of such machine-generated translations: these translations are transferring modern text with their modern genres into a historically "alien" environment. In this way, although the language may fit within the historical register, current MT algorithms are not able to transform the translated text according to its historical stylistics and context, especially for low-resource languages in a *zero-resource setting*, where they make *mistakes on distributionally similar words* [Bap+22]. This very context is actually the main focus of scholars of historical studies aiming to understand old manuscripts and volumes of printed books. There might be an overlap in the language with some regular everyday activities, but, in the domain-specific elaborations, we will have some issues of erroneous translations. Therefore, automatized resource generation should be treated with more care, as it poses only a first approximation of models trained on human annotated data on historical literature.

Hence, as long as we do not have enough resources for a given historical language and its domain, we cannot automatize this procedure with the support of MT tools and must continue with the cost- and time-intensive process of annotating the historical literature word by word along the ladder of NLP tasks.

5.2.3 Watchlist of Low-Resource Languages and Outlook

To ensure further progress in this research direction, we propose the creation of a standardized *watchlist of low-resource languages*, in which each endangered language is classified according to its level of digitization and chance of survival in this modern age (akin to biodiversity research). In that same list, *dead languages* can be included as well, which have large historical treasures (e.g. Sumerian) and can thus be revived, at least virtually, or even to some extent similar to the revival of Hebrew. Our review in Section 2.2 and especially Table 2.1 provide a first indication of such a list from a historical perspective. In a follow-up step, such research endeavors can even enable a revival of ancient civilizations in virtual reality environments.

Overall, our dissertation has opened up various doors and pathways for future research work. It has shown how important resource generation is for the progress of historical NLP. We believe that this is currently the most promising way of ensuring how endangered low-resource languages can cross the next transitional step and

[6] `https://www.bing.com/translator`

become part of the new digital age toward which our society is rapidly moving. Languages are an important aspect of human cultural heritage. Their preservation should be treated the same way as that of nature and its vast biodiversity. We believe that thinking in a given language gives an individual person a unique view of life. In this sense, we think that language can be seen as the "sixth sense" of human beings—a specific lens allowing a specific experience of life.

Still, we believe that our dissertation is just another drop of water in the ocean of knowledge—an ocean if turned into ink for the words describing our world, sooner would the ocean be exhausted than would the words come to an end.

6 German Summary (Deutsche Zusammenfassung)

Einführung

In unserer heutigen Zeit machen wir großen Fortschritt im Bereich der künstlichen Intelligenz (KI). Seit dem Erfolg von neuronalen Netzwerken werden kontinuierlich neue Grenzen entdeckt. Die Entwicklung ist so rasant, dass keine Obergrenzen ersichtlich sind. Im Bereich *Natural Language Processing* (NLP) wird die Mehrheit der Arbeiten für Englisch durchgeführt, eine sogenannte ressourcenreiche (*high-resource*) Sprache, für welche eine große Anzahl an Vorarbeiten und digitale Ressourcen existieren. Dies beschleunigt gewiss den Vorgang der laufenden *Big-Data-getriebenen* Forschung, so wie es aktuell an zahlreichen NLP-Benchmarks (Vergleichsmaßstaben) ersichtlich ist (z.B. SNLI [Bow+15] für *Natural Language Inference*, IIRC [Fer+20] für *Machine Reading Comprehension*, SQUAD [RJL18] für *Question Answering*). Aus Sicht der KI-Forschung ist es in der Tat förderlich, die Forschung anhand einer Sprache fortzusetzen, welche bereits ein hohes Maß an Digitalisierung erfahren hat. Das wird uns sicherlich näher an das (große) Ziel bringen, eine menschenähnliche *starke KI* zu entwickeln. Aus einem sozial-ethischen Betrachtungswinkel ist jedoch dieser vollständige Fokus auf einer Sprache für den Nachteil anderer existierender Sprachen nicht gerecht, angesichts der steigenden Nachfrage an NLP-Modellen für Sprachen nicht-englischem Ursprungs. Folglich ist eine Lücke zwischen modernen, ressourcenreichen und historischen, ressourcenarmen (*low-resource*) Sprachen entstanden.

In vergangenen Zeiten existierten unzählige historische Sprachen, welche für zahlreiche Teile der menschlichen Gesellschaft und ihrer Aktivitäten wichtig waren. Jene Sprachen waren Verkehrssprachen (*Lingua Franca*) für Wissenschaft, Kunst, Handel, und dem alltäglichen Leben. Sprachen wie Altägyptisch, Altgriechisch, klassisches Arabisch, oder vormodernes Deutsch (mit seiner Fraktur Schrift), welche große Mengen an historischer Literatur besitzen, waren und sind auch bis dato relevant für viele (wissenschaftliche) Gemeinschaften und (religiöse) Gesellschaften und beeinflussen sogar jetzt noch maßgebend deren weitere Entwicklungen. Mit dem Untergang jener Zivilisationen (samt ihrer Sprachen) und ihrer Ersetzung durch nachfolgende moderne Zivilisationen wurden nur Teile ihres kulturellen Erbguts fortgetragen. Die

Mehrheit des Erbguts wurde in handgeschriebenen Manuskripten und gedruckten Büchern begraben, von denen nur ein gewisser Anteil unsere heutige Zeit überdauert hat. Diese wichtigen historischen Sprachen mit ihrem immens großen Reichtum verdienen die Aufmerksamkeit der aktuellen, stets wachsenden NLP-Forschung. Um historische Analysen zu ermöglichen, die relevant für unsere moderne Zeit sind, müssen wir diese *vergessenen* Sprachen von der Erfolgswelle des Maschinellen Lernens (ML) profitieren lassen, damit historische Texte modernen wissenschaftlichen Studien zugänglich machen und aus ethischer Sicht einem Gleichgewichtszustand in der NLP-Forschung annähern.

In unserer kumulativen Dissertation erforschen wir den Bereich der *historischen NLP-Forschung*. Wir schließen schrittweise die wachsende Ressourcen- und Leistungslücke durch das Analysieren von zwei ziemlich verschiedenen, ressourcenarmen Sprachen, nämlich vormodernes Deutsch in dem Anwendungsbereich der Biodiversitätsforschung und klassisches Arabisch in dem Anwendungsbereich der theologischen Studien. Wir führen dies anhand der Beispiele von grundlegenden NLP-Aufgaben wie *Sentence Boundary Detection* (SBD) [SA19], *Named Entity Recognition* (NER) [ASM18; Ahm+19; Ahm+22] und *Topic Modeling* (TM) [Ahm+22] durch und legen dabei unseren Fokus insbesondere auf das NER. Indem wir unsere Forschungsarbeit mit dem deutschen NER beginnen und diese mit dem arabischen Pendant abschließen, zeigen wir für unsere ausgewählten Sprachen, dass ein kostenintensiver *Annotationsprozess* für die Digitalisierung von historischer Literatur und ihrer weiterführenden Analyse mithilfe moderner Methoden des NLP notwendig ist. Wir demonstrieren, dass eine Generierung von *Annotationsdaten* für die Überwindung des ressourcenarmen Zustandes einer Sprache essentiell ist, und bieten allgemeine Richtlinien für Forscher mit ähnlichen Unternehmungen.

Auf diese Weise ermöglicht unsere Forschungsarbeit eine automatische Extraktion von historischen Informationen, die bisher tief in den Papiermanuskripten und Bücherhaufen verschiedener Bibliotheken versteckt sind. Durch die Generierung der notwendigen Trainingsdaten für die Analyse von grundlegenden NLP-Aufgaben mit modernen Verfahren des MLs stellen wir einen frei zugänglichen (*open-source*) Goldstandard für die NLP-Fachcommunity bereit und legen damit die Grundbausteine weiterführender zukünftiger Forschungsarbeiten für die Digitalisierung historischer Studien.

In den nächsten Abschnitten folgt die Zusammenfassung der einzelnen Publikationen, die den Hauptteil dieser kumulativen Dissertation bilden. Somit wird uns demonstriert, wie schrittweise, pro Publikation unsere Ziele für solche ressourcenarmen Sprachen erreicht werden.

Ressourcengröße ist wichtig: Verbesserung vom neuronalen Named Entity Recognition mit optimierten umfangreichen Textkorpora (Publikation 1 [ASM18])

In dieser Studie verbessern wir die Leistung von NER mit neuronalen Netzwerken anhand des Beispiels einer ressourcenarmen Sprache wie Deutsch. Dabei übertreffen wir bestehende Normen (*Baselines*) und etablieren eine neue State-of-the-Art auf jedes einzelne *open-source* NER-Datensatz mit einer Verbesserung von bis zu 11% F-Score. Anstatt tiefere und bereitere hybride neuronale Architekturen zu konstruieren, sammeln wir alle verfügbaren Ressourcen und führen eine detaillierte Optimierung und morphologische Vorverarbeitung (u.a. *Lemmatisierung* und *Part-of-Speech Tagging*) durch. Erst dann werden die so vorverarbeiteten rohen Textdaten dem eigentlichen neuronalen Trainingsprozess vorgelegt. Wir testen unsere Vorgehensweise in einem dreifachen, monolinguistischen Versuchsaufbau, nämlich a) einfaches, b) gemeinsames, und c) optimiertes Training; daraufhin analysieren wir die Abhängigkeit der NER-Aufgabe von der Größe der Textkorpora, die verwendet werden, um die grundlegenden Word-Vektoren zu berechnen. Unsere Studie dient zur Vorbereitung des BIOfid-Datensatzes und zeigt, dass für die Endleistung ressourcenarmer Sprachen wie Deutsch neben einer sorgfältigen sprachabhängig Vorverarbeitung die Korpusgröße in der Tat eine zentrale Rolle spielt. Mit unserer Arbeit schließen wir schlussendlich die Leistungs- und Ressourcenlücke zu Englisch, einer der führenden ressourcenreichen Sprache unserer heutigen Zeit.

Deep-EOS: Mehrzweckmodelle aus neuronalen Netzwerken für Satzgrenzenerkennung (Publikation 2 [SA19])

In dieser Studie präsentieren wir drei Mehrzweckmodelle aus neuronalen Netzwerken für die NLP-Aufgabe *Satzgrenzenerkennung* (SBD). Wir berichten über die Experimentenreihe mit *Long Short-Term Memory* (LSTM), *Bidirectional Long Short-Term Memory* (BiLSTM) und *Convolutional Neural Network* (CNN) für SBD. Wir zeigen, dass diese neuronale Netzwerkarchitekturen das populäre Anwendungsframework *OpenNLP*, welches auf die Maximum-Entropie-Methode basiert, in Leistung übertreffen. Auf diese Weise etablieren wir eine neue State-of-the-Art sowohl auf unserer mehrsprachigen Benchmark, welche aus bis zu 12 verschiedenen Sprachen besteht, als auch auf unseren Zero-Shot-Szenarien. Wir kommen zur Schlussfolgerung, dass unsere trainierten Modelle für die Konstruktion eines sprachunabhängigen, robusten SBD-Systems verwendet werden können. Damit dient unsere Studie zur Vorbereitung der Generierung des BIOfid-Datensatzes für das deutsche NER.

BIOfid-Datensatz: Veröffentlichung eines deutschen Goldstandards für Named Entity Recognition in historischer Literatur zur Biodiversität (Publikation 3 [Ahm+19])

Der *Fachinformationsdienst Biodiversitätsforschung* (BIOfid) wurde eingeführt, um wertvolle biologische Daten aus gedruckter Literatur zu mobilisieren, welche seit über 250 Jahren in deutschen Bibliotheken verborgen sind. In diesem Projekt annotieren wir manuell den deutschen Rohtext, der durch einen OCR-Prozess aus historischer wissenschaftlicher Literatur über die Biodiversität von Pflanzen, Vögeln, Motten und Schmetterlingen generiert wurde. Unsere Arbeit ermöglicht eine automatische Extraktion von biologischen Informationen, die bisher tief in der Masse von Papieren und Bücherhaufen begraben waren. Für diesen Zweck generieren wir Trainingsdaten für die NLP-Aufgabe NER und *Taxa Recognition* (TR) in biologischen Dokumenten. Mithilfe dieser Daten trainieren wir eine Anzahl an führenden ML-Tools und schaffen damit einen Goldstandard für TR innerhalb der Biodiversitätsliteratur. Im engeren Sinne führen wir eine praktische Analyse des von uns neu generierten *BIOfid-Datensatzes* durch, indem wir verschiedene Endevaluationen mit aktuellen Sprachmodellen (u.a. leichtgewichtiges BiLSTM, schwergewichtiges ELMo und BERT) durchführen. Dies führt dazu, dass wir eine neue State-of-the-Art für TR mit bis zu 80.23% F-Score etablieren. Mit unserer Studie legen wir somit die Grundbausteine für zukünftige Forschungsarbeiten in dem Bereich der automatischen Analyse von biologischen Texten.

Tafsir-Datensatz: Eine neuartige Benchmark für Named Entity Recognition und Topic Modeling in klassischer arabischer Literatur (Publikation 4 [Ahm+22])

Zahlreiche historische Sprachen, welche früher Verkehrssprachen für Wissenschaft und Kunst waren, verdienen die Aufmerksamkeit der aktuellen NLP-Forschung. In dieser Studie unternehmen wir die ersten Schritte gemäß dieser Forschungsrichtung für das klassische Arabisch am Beispiel der NLP-Aufgaben NER und TM in klassischer arabischer Literatur. Wir annotieren manuell das enzyklopädische Werk *Tafsir Al-Tabari* mit NEs und Topics und generieren damit das *Tafsir-Datensatz* mit über 51.000 Sätzen, die erste umfangreiche, mehrschichtige Benchmark für das klassische Arabisch. Als nächstes analysieren wir unseren neu generierten Datensatz, welchen wie frei (*open-source*) zugänglich machen, mit aktuellen Sprachemodellen (u.a. leichtgewichtiges BiLSTM, Transformer-basiertes MaChAmP) samt einer innovativen Schriftkompressionsmethode. Hierdurch erzielen wir eine neue State-of-the-Art Leistung für unsere Zielaufgabe NER. Wir zeigen zudem, dass TM aus der Perspektive von historischen Topic-Modellen herausfordernd ist, welche für arabische Studien eine zentrale Rolle spielen. Mit unserer interdisziplinären Studie schließen wir die Ressourcen- und Leistungslücke zu ressourcenreichen Sprachen und legen damit die Grundsteine für zukünftige Forschungsarbeiten in dem Bereich der automatischen Analyse von klassischer arabischer Literatur.

Schlussfolgerung

In unserer heutigen Zeit fokussiert sich eine Vielzahl an Forschern aus dem Gebiet des ML-basierten NLP auf tiefe und breite Modelle neuronaler Netzwerke. Je weiter die Forschung voranschreitet, desto tiefer und breiter werden diese Netzwerke. Innerhalb solcher Entwicklungen werden jedoch ressourcenarme Sprachen außer Acht gelassen, was zu wesentlichen wissenschaftlichen aber auch ethischen Problemen führt. In dem Hauptteil unserer Dissertation haben wir schrittweise gesehen, wie durch den Fokus auf die Datenseite des NLP-Fortschritts wir dieser Entwicklung entgegenwirken können und somit ressourcenarme Sprachen vom Prozess der *Ressourcengenerierung und -optimierung* profitieren können. Wir haben gesehen, dass das Sammeln von rohen, unbezeichneten Textdaten (*unlabeled data*) ziemlich wichtig ist. Insbesondere haben wir aber gesehen, dass die Annotationsarbeit (*gold data*) eine entscheidende Rolle dabei spielt, die Ressourcen- und Leistungslücke zwischen modernen, ressourcenreichen und historischen, ressourcenarmen Sprachen zu schließen. Auf diese Weise kann eine automatische Analyse von historischer Literatur durchgeführt werden.

Diese Dissertation leistete einen Beitrag zu zahlreichen Bereichen der Wissenschaft. Unsere Arbeit hat gezeigt, ob aktuelle ML-Methoden für die Verarbeitung großer Sammlungen historischer Texte sowohl zur Biodiversität als auch zur Theologie geeignet sind. Für diese jeweiligen Bereiche wurde die Entwicklung von frei verfügbaren Datensätzen und Online-Suchwerkzeugen initiiert, die es Fachwissenschaftlern ohne technischen Hintergrund ermöglichen, ihre langwierigen Forschungsanfragen durchzuführen. Dies ist sowohl der Fall für die historische Forschung zur Biodiversität mit der Online-Plattform *BIOfid*[1], als auch zur Religionswissenschaften mit der Online-Plattform *Linked Open Tafsir*[2]. Damit ermöglicht dies den Zugang zu wertvollen, bisher unberührten Textdaten in solch eine Weise, dass es womöglich zu neuen Durchbrüchen in der Forschung sowohl zum Klimawandel als auch zur theologischen Literaturwissenschaft führen kann.

Darüber hinaus hat unsere Arbeit die ML-Methoden auf die spezifischen Anforderungen ausgedehnt, die im Kontext der wichtigen, aber trotzdem ressourcenarmen

[1] https://www.biofid.de/

[2] https://www.tafsirtabari.com/

Sprachen wie Deutsch und Arabisch ergeben. Wir haben gesehen, dass trotz der kulturellen, historischen und sprachlichen Unterschiede unserer beiden ressourcenarmen Zielsprachen aus den indogermanischen und afroasiatischen Sprachfamilien ein gemeinsames Transformationsverfahren zu vergleichbaren Verbesserungen führt. Mit dieser Forschungsarbeit haben wir Schritte unternommen, um der englisch dominierten Landschaft der NLP-Forschung entgegenzuwirken. Die Verwendung der englischen Sprache als Hauptmedium für die Forschung hat zweifellos vielzählige Vorteile. Es ist die Verkehrssprache der modernen Ära. Zudem ist es linguistisch gesehen eine analytische Sprache, folglich besitzt es eine geringe Anzahl an grammatikalischen Flexionen [McA92], was sie für die ML-Methoden sicherlich besser geeignet macht. Jedoch ist es gut möglich, dass wir neue Grenzen der NLP-Forschung überschreiten, falls wir mit anderen, noch komplexeren Sprachen (wie Latein, klassisches Arabisch, Sanskrit, und sogar Esperanto) arbeiten, welche neue Blickwinkel auf alte Forschungsfragen eröffnen können. Alles in allem hat unsere Arbeit zum Fortschritt der künstlichen Intelligenz zum menschenähnlichen Niveau beigetragen und dabei neue Grenzen im aufstrebenden Bereich der historischen NLP-Forschung erreicht und wohlmöglich auch überschritten.

Trotz allem glauben wir daran, dass unsere Dissertation nur ein Tropfen Wasser in dem Ozean des Wissens ist – ein Ozean falls zur Tinte verwandelt für die Worte unserer Welt, würde der Ozean eher ausgeschöpft sein, als würden die Worte ein Ende finden.

Bibliography

[Abd+22] Nora Abdelmageed, Felicitas Löffler, Leila Feddoul, et al. „BiodivNERE: Gold standard corpora for named entity recognition and relation extraction in the biodiversity domain“. In: *Biodiversity Data Journal* 10 (2022), e89481 (cit. on p. 14).

[Ahm+19] Sajawel Ahmed, Manuel Stoeckel, Christine Driller, Adrian Pachzelt, and Alexander Mehler. „BIOfid Dataset: Publishing a German Gold Standard for Named Entity Recognition in Historical Biodiversity Literature“. In: *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 871–880 (cit. on pp. 2, 3, 7, 12, 14, 74, 76).

[Ahm+22] Sajawel Ahmed, Rob van der Goot, Misbahur Rehman, et al. „Tafsir Dataset: A Novel Multi-Task Benchmark for Named Entity Recognition and Topic Modeling in Classical Arabic Literature“. In: *Proceedings of the 29th International Conference on Computational Linguistics*. Gyeongju, Republic of Korea, Oct. 2022, pp. 3753–3768 (cit. on pp. 2, 3, 7, 10, 12, 14, 64, 69, 74, 76).

[ASM18] Sajawel Ahmed, Manuel Stoeckel, and Alexander Mehler. „Resource-Size matters: Improving Neural Named Entity Recognition with Optimized Large Corpora“. In: *Proceedings of the 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. © 2018 IEEE. Reprinted, with permission, from above mentioned authors, and IEEE publication title. Orlando, Florida, USA, Dec. 2018 (cit. on pp. 2, 3, 7, 12, 64, 74, 75).

[Ban+21] Rachit Bansal, Himanshu Choudhary, Ravneet Punia, et al. „How Low is Too Low? A Computational Perspective on Extremely Low-Resource Languages“. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*. Online: Association for Computational Linguistics, Aug. 2021, pp. 44–59 (cit. on p. 13).

[Bap+22] Ankur Bapna, Isaac Caswell, Julia Kreutzer, et al. *Building Machine Translation Systems for the Next Thousand Languages*. Tech. rep. Google Research, 2022 (cit. on pp. 69, 70).

[BB20] David Bamman and Patrick J Burns. „Latin bert: A contextual language model for classical philology“. In: *arXiv preprint arXiv:2009.10053* (2020) (cit. on p. 13).

[Ben+03] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. „A Neural Probabilistic Language Model“. In: *J. Mach. Learn. Res.* 3.null (Mar. 2003), pp. 1137–1155 (cit. on p. 10).

[Boj+17] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. „Enriching word vectors with subword information“. In: *Transactions of the association for computational linguistics* 5 (2017), pp. 135–146 (cit. on p. 10).

[Bow+15] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. „A large annotated corpus for learning natural language inference“. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, Sept. 2015, pp. 632–642 (cit. on pp. 1, 73).

[BRB07] Yassine Benajiba, Paolo Rosso, and Jose Miguel Benediruiz. „Anersys: An arabic named entity recognition system based on maximum entropy“. In: *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer. 2007, pp. 143–153 (cit. on p. 14).

[Bro+20] Tom Brown, Benjamin Mann, Nick Ryder, et al. „Language Models are Few-Shot Learners“. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901 (cit. on pp. 9, 12).

[CLS14] Xinxiong Chen, Zhiyuan Liu, and Maosong Sun. „A Unified Model for Word Sense Representation and Disambiguation“. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1025–1035 (cit. on p. 11).

[Con+20] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, et al. „Unsupervised Cross-lingual Representation Learning at Scale“. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 8440–8451 (cit. on p. 12).

[Dev+19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. „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)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186 (cit. on pp. 4, 9, 11, 12).

[DR19] Kshitij Dwivedi and Gemma Roig. „Representation similarity analysis for efficient task taxonomy & transfer learning“. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 12387–12396 (cit. on p. 11).

[Erd+19] Alexander Erdmann, David Joseph Wrisley, Benjamin Allen, et al. „Practical, Efficient, and Customizable Active Learning for Named Entity Recognition in the Digital Humanities“. 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)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 2223–2234 (cit. on p. 13).

[Fer+20] James Ferguson, Matt Gardner, Hannaneh Hajishirzi, Tushar Khot, and Pradeep Dasigi. „IIRC: A Dataset of Incomplete Information Reading Comprehension Questions“. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 1137–1147 (cit. on pp. 1, 73).

[Gle+09] Rüdiger Gleim, Ulli Waltinger, Alexandra Ernst, et al. „The eHumanities Desktop – An Online System for Corpus Management and Analysis in Support of Computing in the Humanities“. In: *Proceedings of the Demonstrations Session of the 12th Conference of the European Chapter of the Association for Computational Linguistics EACL 2009, 30 March – 3 April, Athens*. 2009 (cit. on p. 13).

[GNB10] Martin Gerner, Goran Nenadic, and Casey M Bergman. „LINNAEUS: a species name identification system for biomedical literature“. In: *BMC bioinformatics* 11.1 (2010), pp. 1–17 (cit. on p. 14).

[HS97] Sepp Hochreiter and Jürgen Schmidhuber. „Long short-term memory“. In: *Neural computation* 9.8 (1997), pp. 1735–1780 (cit. on p. 10).

[KM16] Alexandros Komninos and Suresh Manandhar. „Dependency Based Embeddings for Sentence Classification Tasks.“ In: *HLT-NAACL*. 2016, pp. 1490–1500 (cit. on p. 10).

[Lam+16] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. „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: Association for Computational Linguistics, 2016, pp. 260–270 (cit. on p. 4).

[LD21] Dieuwertje Luitse and Wiebke Denkena. „The great transformer: Examining the role of large language models in the political economy of AI“. In: *Big Data & Society* 8.2 (2021), p. 20539517211047734 (cit. on p. 11).

[LG14] Omer Levy and Yoav Goldberg. „Dependency-Based Word Embeddings.“ In: *ACL (2)*. 2014, pp. 302–308 (cit. on p. 10).

[Li+20] Yang Li, Ranjitha Kumar, Walter S. Lasecki, and Otmar Hilliges. „Artificial Intelligence for HCI: A Modern Approach“. In: *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI EA ’20. Honolulu, HI, USA: Association for Computing Machinery, 2020, pp. 1–8 (cit. on p. 4).

[Lüc+21] Andy Lücking, Christine Driller, Manuel Stoeckel, et al. „Multiple Annotation for Biodiversity: Developing an annotation framework among biology, linguistics and text technology“. In: *Language Resources and Evaluation* (2021). Ed. by Nancy Ide and Nicoletta Calzolari (cit. on p. 14).

[McA92] Tom McArthur. *Oxford companion to the English language*. Oxford University Press, 1992, p. 64 (cit. on pp. 64, 68, 78).

[McC+17] Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. „Learned in Translation: Contextualized Word Vectors“. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS’17. Long Beach, California, USA: Curran Associates Inc., 2017, pp. 6297–6308 (cit. on p. 11).

[Meh+10] Alexander Mehler, Rüdiger Gleim, Ulli Waltinger, and Nils Diewald. „Time Series of Linguistic Networks by Example of the Patrologia Latina“. In: *Proceedings of INFORMATIK 2010: Service Science, September 27 - October 01, 2010, Leipzig*. Ed. by Klaus-Peter Fähnrich and Bogdan Franczyk. Vol. 2. Lecture Notes in Informatics. GI, 2010, pp. 609–616 (cit. on p. 13).

[Meh+15] Alexander Mehler, Tim vor der Brück, Rüdiger Gleim, and Tim Geelhaar. „Towards a Network Model of the Coreness of Texts: An Experiment in Classifying Latin Texts using the TTLab Latin Tagger“. In: *Text Mining: From Ontology Learning to Automated text Processing Applications*. Ed. by Chris Biemann and Alexander Mehler. Theory and Applications of Natural Language Processing. Berlin/New York: Springer, 2015, pp. 87–112 (cit. on p. 13).

[Meh+20] Alexander Mehler, Bernhard Jussen, Tim Geelhaar, et al. „The Frankfurt Latin Lexicon. From Morphological Expansion and Word Embeddings to SemioGraphs“. In: *Studi e Saggi Linguistici* 58.1 (2020), pp. 121–155 (cit. on p. 13).

[Mik+13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. „Distributed representations of words and phrases and their compositionality“. In: *Advances in neural information processing systems*. 2013, pp. 3111–3119 (cit. on pp. 4, 10).

[Moh+12] Behrang Mohit, Nathan Schneider, Rishav Bhowmick, Kemal Oflazer, and Noah A Smith. „Recall-oriented learning of named entities in Arabic Wikipedia“. In: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. 2012, pp. 162–173 (cit. on p. 14).

[MRS18] Matthew Thomas Miller, Maxim G. Romanov, and Sarah Bowen Savant. „Digitizing the Textual Heritage of the Premodern Islamicate World: Principles and Plans“. In: *International Journal of Middle East Studies* 50.1 (2018), pp. 103–109 (cit. on p. 64).

[MS21] Muhammad Majadly and Tomer Sagi. „Dynamic Ensembles in Named Entity Recognition for Historical Arabic Texts“. In: *Proceedings of the Sixth Arabic Natural Language Processing Workshop*. Kyiv, Ukraine (Virtual): Association for Computational Linguistics, Apr. 2021, pp. 115–125 (cit. on p. 14).

[NGA19] Nhung T.H. Nguyen, Roselyn S. Gabud, and Sophia Ananiadou. „COPIOUS: A gold standard corpus of named entities towards extracting species occurrence from biodiversity literature“. In: *Biodiversity Data Journal* 7 (2019), e29626 (cit. on p. 14).

[Paf+13] Evangelos Pafilis, Sune P. Frankild, Lucia Fanini, et al. „The SPECIES and ORGANISMS Resources for Fast and Accurate Identification of Taxonomic Names in Text“. In: *PLOS ONE* 8 (June 2013), pp. 1–6 (cit. on p. 14).

[Pap+02] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. „Bleu: a Method for Automatic Evaluation of Machine Translation“. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, July 2002, pp. 311–318 (cit. on p. 70).

[Pet+18] Matthew E. Peters, Mark Neumann, Mohit Iyyer, et al. „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: Association for Computational Linguistics, June 2018, pp. 2227–2237 (cit. on pp. 11, 12).

[Phi90] Lawrence Philips. „Hanging on the metaphone“. In: *Computer Language*. Vol. 7. 1990, pp. 39–43 (cit. on p. 69).

[PSM14] Jeffrey Pennington, Richard Socher, and Christopher Manning. „GloVe: Global Vectors for Word Representation“. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543 (cit. on p. 10).

[RJL18] Pranav Rajpurkar, Robin Jia, and Percy Liang. „Know What You Don't Know: Unanswerable Questions for SQuAD“. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 784–789 (cit. on pp. 1, 73).

[RKR20] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. „A Primer in BERTology: What We Know About How BERT Works“. In: *Transactions of the Association for Computational Linguistics* 8 (2020), pp. 842–866 (cit. on p. 11).

[Rud+21] Sebastian Ruder, Noah Constant, Jan Botha, et al. „XTREME-R: Towards More Challenging and Nuanced Multilingual Evaluation“. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 10215–10245 (cit. on p. 15).

[SA19] Stefan Schweter and Sajawel Ahmed. „Deep-EOS: General-Purpose Neural Networks for Sentence Boundary Detection“. In: *Proceedings of the 15th Conference on Natural Language Processing (KONVENS)*. Erlangen, Germany, 2019 (cit. on pp. 2, 3, 7, 74, 75).

[San+21] Jivnesh Sandhan, Om Adideva, Digumarthi Komal, Laxmidhar Behera, and Pawan Goyal. „Evaluating Neural Word Embeddings for Sanskrit“. In: *arXiv preprint arXiv:2104.00270* (2021) (cit. on p. 13).

[Sch15] Roland Schäfer. „Processing and querying large web corpora with the COW14 architecture“. In: *Proceedings of Challenges in the Management of Large Corpora 3 (CMLC-3)*. Ed. by Piotr Baski, Hanno Biber, Evelyn Breiteneder, et al. UCREL. Lancaster: IDS, 2015 (cit. on p. 64).

[Shm+22] Avi Shmidman, Joshua Guedalia, Shaltiel Shmidman, et al. „Introducing BEREL: BERT Embeddings for Rabbinic-Encoded Language“. In: *arXiv preprint arXiv: 2208.01875* (2022) (cit. on p. 13).

[SM17] Mohamad Bagher Sajadi and Behrooz Minaei. „Arabic named entity recognition using boosting method“. In: *2017 Artificial Intelligence and Signal Processing Conference (AISP)*. IEEE. 2017, pp. 281–288 (cit. on p. 14).

[SRL21] Pranaydeep Singh, Gorik Rutten, and Els Lefever. „A Pilot Study for BERT Language Modelling and Morphological Analysis for Ancient and Medieval Greek“. In: *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Punta Cana, Dominican Republic (online): Association for Computational Linguistics, Nov. 2021, pp. 128–137 (cit. on p. 13).

[Sto+20] Manuel Stoeckel, Alexander Henlein, Wahed Hemati, and Alexander Mehler. „Voting for POS tagging of Latin texts: Using the flair of FLAIR to better Ensemble Classifiers by Example of Latin“. In: *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*. Marseille, France: European Language Resources Association (ELRA), May 2020, pp. 130–135 (cit. on p. 13).

[SZ18] Ramzi Esmail Salah and Lailatul Qadri Binti Zakaria. „Building the classical Arabic named entity recognition corpus (CANERCorpus)“. In: *2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP)*. IEEE. 2018, pp. 1–8 (cit. on p. 14).

[TD03] Erik F. Tjong Kim Sang and Fien De Meulder. „Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition“. In: *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*. 2003, pp. 142–147 (cit. on p. 14).

[Tia+14] Fei Tian, Hanjun Dai, Jiang Bian, et al. „A probabilistic model for learning multi-prototype word embeddings“. In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. 2014, pp. 151–160 (cit. on p. 11).

[Tia+21] Huishuang Tian, Kexin Yang, Dayiheng Liu, and Jiancheng Lv. „Anchibert: a pre-trained model for ancient Chinese language understanding and generation“. In: *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2021, pp. 1–8 (cit. on p. 13).

[Vas+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. „Attention is All you Need“. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. Von Luxburg, S. Bengio, et al. Vol. 30. Curran Associates, Inc., 2017 (cit. on p. 9).

[Wan+20] Yuxuan Wang, Yutai Hou, Wanxiang Che, and Ting Liu. „From static to dynamic word representations: a survey“. In: *International Journal of Machine Learning and Cybernetics* 11.7 (2020), pp. 1611–1630 (cit. on p. 11).

[Wan+21] Xueting Wang, Fang Miao, Huixin Liu, Guoting Zhang, and Libiao Jin. „Joint Extraction of Entities and Relations from Ancient Chinese Medical Literature“. In: *2021 International Conference on Culture-oriented Science & Technology (ICCST)*. IEEE. 2021, pp. 369–372 (cit. on p. 13).

[WLH22] Guanghai Wang, Yudong Liu, and James Hearne. „Few-shot Learning for Sumerian Named Entity Recognition“. In: *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*. 2022, pp. 136–145 (cit. on p. 13).

[Wol+20] Thomas Wolf, Lysandre Debut, Victor Sanh, et al. „Transformers: State-of-the-Art Natural Language Processing“. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45 (cit. on p. 12).

[Wu+16] Yonghui Wu, Mike Schuster, Zhifeng Chen, et al. „Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation“. In: *CoRR* abs/1609.08144 (2016) (cit. on p. 69).

[Yan+19] Zhilin Yang, Zihang Dai, Yiming Yang, et al. „XLNet: Generalized Autoregressive Pretraining for Language Understanding“. In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, et al. Vol. 32. Curran Associates, Inc., 2019 (cit. on pp. 9, 12).

[YB18] Vikas Yadav and Steven Bethard. „A Survey on Recent Advances in Named Entity Recognition from Deep Learning models“. In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 2145–2158 (cit. on p. 10).

[Zic+21] Roberto V Zicari, John Brodersen, James Brusseau, et al. „Z-Inspection®: a process to assess trustworthy AI“. In: *IEEE Transactions on Technology and Society* 2.2 (2021), pp. 83–97 (cit. on p. 4).

[ZMT20] Amir Zeldes, Lance Martin, and Sichang Tu. „Exhaustive Entity Recognition for Coptic: Challenges and Solutions“. In: *Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Online: International Committee on Computational Linguistics, Dec. 2020, pp. 19–28 (cit. on p. 13).

List of Figures

List of Tables

Appendix A

A.1 Publications

This cumulative dissertation was based on the following study papers:

- Sajawel Ahmed, Manuel Stoeckel and Alexander Mehler. *Resource-Size matters: Improving Neural Named Entity Recognition with Optimized Large Corpora.* In: Proceedings of the 17th IEEE International Conference on Machine Learning and Applications (ICMLA). Orlando, Florida, USA. 2018.

- Stefan Schweter and Sajawel Ahmed. *Deep-EOS: General-Purpose Neural Networks for Sentence Boundary Detection.* In: Proceedings of the 15th Conference on Natural Language Processing (KONVENS). Erlangen, Germany. 2019.

- Sajawel Ahmed, Manuel Stoeckel, Christine Driller, Adrian Pachzelt and Alexander Mehler. *BIOfid Dataset: Publishing a German Gold Standard for Named Entity Recognition in Historical Biodiversity Literature.* In: Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL). Hong Kong, China. 2019.

- Sajawel Ahmed, Rob van der Goot, Misbahur Rehman, Carl Kruse, Ömer Özsoy, Alexander Mehler and Gemma Roig. *A Novel Multi-Task Benchmark for Named Entity Recognition and Topic Modeling in Classical Arabic Literature.* In: Proceedings of the 29th International Conference on Computational Linguistics (COLING). Gyeongju, Republic of Korea. 2022.

Additionally, I was also co-author of the following study papers while pursuing my doctorate. These papers were not directly incorporated into the dissertation, although content-wise they provide further details and can be examined by those interested in a broader picture of the study papers and their project environments.

- Claus Weiland, Christine Driller, Markus Koch, Marco Schmidt, Guiseppe Abrami, Sajawel Ahmed, Alexander Mehler, Adrian Pachzelt, Gerwin Kasperek, Angela Hausinger and Thomas Hörnschemeyer. *BioFID, a platform to enhance* *accessibility of biodiversity data.* In: Proceedings of the 10th International Conference on Ecological Informatics. 2018.

- Christine Driller, Markus Koch, Marco Schmidt, Claus Weiland, Thomas Hörnschemeyer, Thomas Hickler, Guiseppe Abrami, Sajawel Ahmed, Rüdiger Gleim, Wahed Hemati, Tolga Uslu, Alexander Mehler, Adrian Pachzelt, Jashar Rexhepi, Thomas Risse, Janina Schuster, Gerwin Kasperek and Angela Hausinger. *Workflow and Current Achievements of BIOfid, an Information Service Mobilizing Biodiversity Data from Literature Sources*. In: Biodiversity Information Science and Standards, vol. 2, p. e25876. 2018.

- Manuel Stoeckel, Sajawel Ahmed, and Alexander Mehler. *SenseFitting: Sense Level Semantic Specialization of Word Embeddings for Word Sense Disambiguation.* In: arXiv preprint arXiv:1907.13237. 2019.

- Sajawel Ahmed, Misbahur Rehman, Edin Mahmutovic, Joschua Tischlik , Carl Kruse and Ömer Özsoy. *Linked Open Tafsir – Rekonstruktion der Entstehungsdynamik(en) des Korans* In: 8. Jahrestagung des Verbandes Digital Humanities im deutschsprachigen Raum (DHd). 2022.

A.2 Co-Advised Thesis and Seminar Work

During the course of my doctorate, I also had the privilege to co-advice the following thesis and seminar works:

- Manuel Stoeckel. *Deep-Learning basierte Methoden für die automatische Disambiguierung am Beispiel von Wiki-basierten Systemen.* Bachelor Thesis, Goethe University Frankfurt, 2018.

- Carl-Sven Kruse. *Digital Humanities and Islamic Studies: Script Analysis for Classical Arabic Texts.* Seminar Work, Goethe University Frankfurt, 2022.

- Anja Meyer. *Digital Humanities: Wie kann uns maschinelles Lernen bei der historisch-kritischen Quellenanalyse für das klassische Arabisch unterstützen?* Seminar Work, Johannes Gutenberg University Mainz, 2022

A.3 Statement of Author Contribution

In the following, details of the specific contributions are provided for all study papers which are part of this cumulative dissertation.

- Sajawel Ahmed, Manuel Stoeckel and Alexander Mehler. *Resource-Size matters: Improving Neural Named Entity Recognition with Optimized Large Corpora.* In: Proceedings of the 17th IEEE International Conference on Machine Learning and Applications (ICMLA). Orlando, Florida, USA. 2018.

  Research Idea: Sajawel Ahmed, Manuel Stoeckel and Alexander Mehler; Data Analysis: Sajawel Ahme and Manuel Stoeckel; Result Interpretation: Sajawel Ahmed; Writing: Sajawel Ahmed and Alexander Mehler

  Contribution percentages: Sajawel Ahmed 65%, Manuel Stoeckel 5%, Alexander Mehler 30%

- Stefan Schweter and Sajawel Ahmed. *Deep-EOS: General-Purpose Neural Networks for Sentence Boundary Detection.* In: Proceedings of the 15th Conference on Natural Language Processing (KONVENS). Erlangen, Germany. 2019.

  Research Idea: Stefan Schweter and Sajawel Ahmed; Data Analysis: Stefan Schweter and Sajawel Ahmed; Result Interpretation: Stefan Schweter and Sajawel Ahmed; Writing: Stefan Schweter and Sajawel Ahmed

  Contribution percentages: Stefan Schweter 50%, Sajawel Ahmed 50%

- Sajawel Ahmed, Manuel Stoeckel, Christine Driller, Adrian Pachzelt and Alexander Mehler. *BIOfid Dataset: Publishing a German Gold Standard for Named Entity Recognition in Historical Biodiversity Literature.* In: Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL). Hong Kong, China. 2019.

  Research Idea: Sajawel Ahmed and Alexander Mehler; Design of Annotation Process: Sajawel Ahmed; (Annotation) Data Analysis: Sajawel Ahmed, Manuel Stoeckel; Result Interpretation: Sajawel Ahmed; Writing: Sajawel Ahmed, Manuel Stoeckel, Christine Driller, Adrian Pachzelt and Alexander Mehler.

  Contribution percentages: Sajawel Ahmed 60%, Manuel Stoeckel 15%, Christine Driller 5%, Adrian Pachzelt 5%, Alexander Mehler 15%

- Sajawel Ahmed, Rob van der Goot, Misbahur Rehman, Carl Kruse, Ömer Özsoy, Alexander Mehler and Gemma Roig. *A Novel Multi-Task Benchmark for Named Entity Recognition and Topic Modeling in Classical Arabic Literature.* In: Proceedings of the 29th International Conference on Computational Linguistics (COLING). Gyeongju, Republic of Korea. 2022.

  Research Idea: Sajawel Ahmed, Ömer Özsoy, Gemma Roig; Design of Annotation Process: Sajawel Ahmed and Misbahur Rehman; (Annotation) Data Analysis: Sajawel Ahmed, Rob van der Goot, Carl Kruse; Result Interpretation: Sajawel Ahmed, Rob van der Goot, Gemma Roig; Writing: Sajawel Ahmed, Rob van der Goot, Alexander Mehler, Gemma Roig

  Contribution percentages: Sajawel Ahmed 60%, Rob van der Goot 10%, Misbahur Rehman 5%, Carl Kruse 2%, Ömer Özsoy 3%, Alexander Mehler 5% and Gemma Roig 15%

A.4 Author's Declaration

I herewith declare that I have produced my doctoral dissertation on the topic of

Data Driven Enrichment of Historical Low-Resource Languages for Foundational NLP Tasks and their Neural Network Models

independently and using only the tools indicated therein. In particular, all references borrowed from external sources are clearly acknowledged and identified. I confirm that I have respected the principles of good scientific practice and have not made use of the services of any commercial agency in respect of my doctorate.

Frankfurt, Germany, 2023

Sajawel Ahmed