Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 4458–4465
Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

Czech Historical Named Entity Corpus v 1.0

Helena Hubková♣, Pavel Král♣, Eva Pettersson♠
♣Department of Computer Science and Engineering, Faculty of Applied Sciences, University of West Bohemia,
Univerzitní 2732/8, 301 00 Pilsen, Czech Republic
♠Dept. of Linguistics and Philology, Uppsala University
P.O. Box 256, SE-751 05 Uppsala, Sweden
♣{hhubkova, pkral}@kiv.zcu.cz, ♠eva.pettersson@lingfil.uu.se

Abstract

As the number of digitized archival documents increases very rapidly, named entity recognition (NER) in historical documents has become very important for information extraction and data mining. For this task an annotated corpus is needed, which has up to now been missing for Czech. In this paper we present a new annotated data collection for historical NER, composed of Czech historical newspapers. This corpus is freely available for research purposes at `http://chnec.kiv.zcu.cz/`. For this corpus, we have defined relevant domain-specific named entity types and created an annotation manual for corpus labelling. We further conducted some experiments on this corpus using recurrent neural networks in order to in order to show baseline results on this dataset. We experimented with randomly initialized embeddings and static and dynamic fastText word embeddings. We achieved 0.73 F1 score with a bidirectional LSTM model using static fastText embeddings.

Keywords: Historical Czech, Historical Named Entity Corpus, LSTM, Named Entity Recognition, Neural Networks

1. Introduction

Named entity recognition (NER) is a fundamental task in natural language processing (NLP). As the amount of digitized archival material has increased rapidly during the last few decades, NER has become an important step for information extraction in the field of historical document analysis. However, a lack of annotated historical data for NER is an obstacle to research in this area. Therefore this paper introduces a new annotated data collection dedicated to historical NER with some experiments using methods based on neural networks. The experiments are conducted to show baseline results on this dataset. The corpus is freely available for research purposes at `http://chnec.kiv.zcu.cz/`. We also plan to submit this corpus to be a part of the Language Research Infrastructure of LINDAT/CLARIN project.

This research is carried out within the framework of the project *Modern Access to Historical Sources*, presented in the Porta Fontium portal.[1] One goal of this project is to enable intelligent full-text access to the printed historical documents from the Czech-Bavarian border region. Accordingly, our original data sources to create the corpus are scanned texts from a Czech historical newspaper *Posel od Čerchova* from the second half of the 19^{th} century.

The scanned materials were digitized by optical character recognition (OCR). Then, we defined specific named entities based on the project purpose itself in combination with named-entity types from Ševčíková et al. (2007b). We also created an annotation manual for the final Czech historical named entity corpus.

For our experiments, we created a bidirectional LSTM model for sequence-labelling inspired by Chiu and Nichols (2016). We compared the results of different architectures of the model, i.e., long short-term memory (LSTM) network and bidirectional LSTM (BiLSTM).

We also provide a qualitative analysis of the tagged output text to explore what linguistic phenomena in the historical input data caused problems for automatic NE detection and classification.

2. Related Work

The issue of NER in historical texts has been previously described by several researchers. Grover et al. (2008) built a rule-based NER system for recognizing names of places and persons in digitized records of British parliamentary proceedings from two different periods, the late 17^{th} and early 19^{th} centuries. They focused on issues caused by the nature of historical texts (e.g., a high level of variance in the use of word-initial upper-case letters) as well as issues connected to the use of OCR technology. They found that recognition of personal names achieved better results than recognition of place names. In other words, finding patterns for recognising personal names was easier and more resistant to OCR errors. On the other hand, they also described other problems caused by OCR errors: wrong interpretation of layout, wrong division of tokens and wrong division of paragraphs (in the middle of a token). They reached an F1 score of 71.81% for the period 1814–1817 and 70.35% for the period 1685–1691.

Packer et al. (2010) experimented with recognition of personal names using noisy OCRed data. They tried three different approaches and evaluated the output against hand-labelled test data. They showed that the character-level errors in OCRed data have small impact on NER in comparison to word order errors.

[1] http://www.portafontium.eu/

Rodriguez et al. (2012) evaluated four different tools for NER in historical texts: a) OpenNLP , b) Stanford NER (Finkel et al., 2005), c) AlchemyAPI , d) OpenCalais. They used the Wiener Library data set (4,415 words) and the King College London data set (16,982 words) as input and defined three NE types: *person*, *location* and *organization*. They showed that the Stanford NER system had the overall best performance, especially in case of person and location entities. Similarly, Alchemy API worked best for the NE type organization for manually corrected text, whereas OpenNLP showed the lowest overall accuracy.

Mac Kim and Cassidy (2015) applied the Stanford NER system to the 155 million OCRed articles from historical Australian newspapers to recognize the NE types *person*, *location* and *organization* and they showed how the data can be exploited using a clustering method.

Neudecker (2016) created an open corpus for NER in Dutch (182,483 tokens), French (207,000) and German (96,735) based on OCRed historical newspapers. The work was included in the Europeana Newspapers project,[2] and they used the Stanford NER system for preprocessing German data, whereas the actual NEs were annotated manually. They distinguished the NE types *person*, *location* and *organization* in the corpus.

Moreover, NER in Czech has a quite long tradition in terms of data for the contemporary Czech language. Ševčíková et al. (2007a) introduced two-level classification and used that for manually annotating 11,000 NEs. Based on that, they developed a Czech NE tagger. They distinguished between NE *span recognition* (all NEs are found but the type is not relevant), NE *supertype recognition* (all NEs are found and supertype - first-level - is correct) and NE *type recognition* (all NEs are found and both supertype and type - second-level - are correct). They evaluated the tagger using precision, recall and F-measure metrics. For all NE instances, they got precision 74%, recall 54% and F-measure 62% in case of correct type, and precision 81%, recall 59% and F-measure 68% in case of correct supertype and finally, precision 88%, recall 64% and F-measure 75% in case of correct span.

Similarly, Kravalová and Žabokrtský (2009) presented the Czech NE corpus (CNEC) which used the two-level classification scheme. It consists of around 6,000 sentences (150,022 words). They also used a Support Vector Machine classification approach for training and evaluating data for NER. They distinguished NEs according to Ševčíková et al. (2007a). They got precision 75%, recall 62% and F-measure 68% for type recognition (span and type). In case of correct span and supertype, they achieved precision 75%, recall 67% and F-measure 71%. However, span recognition itself achieved a precision of 84%, a recall of 70% and an F-measure of 76%.

Also, Král (2011) created a NER system for the Czech News Agency to evaluate different features for NER to find an "optimal" set of features. They classified the system using Conditional Random Fields (CRFs) and the evaluation was performed based on the Czech NER corpus (Kravalová and Žabokrtský, 2009). They achieved an F-measure of 58% with the best feature set.

Straková et al. (2013) built a NER system based on a Maximum Entropy Markov Model and a Viterbi algorithm, and evaluated it for Czech and English. They achieved an f-measure of 82.82% for Czech using the Czech Named Entity Corpus (version 1.0) and an F-measure of 89.16% for English using the CoNLL-2003 data set.[3]

Similarly, Straková et al. (2014) presented two open-source taggers: NER tagger *NameTag* and *MorphoDiTa* (Morphological Dictionary and Tagger) for morphological analysis. Both tools are specifically designed for inflective languages including Czech.

Experiments with neural networks are quite common in NER research nowadays. For example, Collobert et al. (2011) presented the unified multilayer convolutional neural network (CNN) model with a learning algorithm which can be used in various NLP tasks including NER. They experimented with training data which were mostly unlabelled and not optimized for each NLP task. For the NER task, they achieved an F1 score of 81.47% for random category (embedding vectors are initialized randomly) and 89.59% for Senna category (using Senna word-embeddings).

Huang et al. (2018) compared different Long Short-Term Memory (LSTM) approaches for sequence tagging. They worked with bidirectional LSTM (BI-LSTM) networks, LSTM with a Conditional Random Field (CRF) layer and bidirectional LSTM with a CRF layer (BI-LSTM-CRF). They showed that using both past and future input in the bidirectional component of BI-LSTM-CRF is efficient and, also, that the CRF layer of the model helps by using sentence level tag information. The system achieved state-of-the-art accuracy results in terms of part-of-speech tagging, chunking and NER data sets. They compare their results with Collobert et al. (2011) and they got an F1 score of 84.26% for random initialized vector embeddings and 90.10% for NER using Senna word-embeddings.

Also, Chiu and Nichols (2016) built a hybrid bidirectional LSTM and CNN model which automatically detects character-level and word-level features. They showed that the system has similar performance to the CoNLL-2003 data set and, moreover, the performance is 2.13 F1 points better than previous research using OntoNotes 5.0. They achieved an F1 score of 91.62% for CoNLL-2003 data and 86.28% for OntoNotes.

Finally, Lample et al. (2016) introduced two neural models - bidirectional LSTM CRF and a transition-based model using shift-reduce parsers. For their experiments, they used character-based word representations based on the supervised corpus, and unsupervised word representations based on the unannotated

[2]http://www.europeana-newspapers.eu/

[3]https://www.clips.uantwerpen.be/conll2003/ner/

corpora. Both models achieved better results than previous research including models using external resources (e.g. gazetteers). Concretely, the LSTM CRF model reached an F1 score of 90.94% for English NER, an F1 score of 78.76% for German, 81.74% for Dutch and 85.75% for Spanish using labelled training external data. In the case of English NER, the LSTM CRF model which was pre-trained by word embeddings, includes character-based modeling of words and dropout rate, achieved and F1 score of 90.94%.

3. Corpus Description

We used the Czech historical newspapers *Posel od Čerchova* as the text source for our corpus. These newspapers were published during the period 1872–1935. However, only scans of issues up to and including 1900 are available on the Porta Fontium portal so far. To begin with, we chose 32 issues from 1872 for further annotations. This number of issues guaranteed more than 70,000 tokens for the final corpus.

3.1. Historical Data vs. Contemporary Czech

Since our data were originally published in 1872, they differ especially in vocabulary, word forms, spelling and word order in comparison to contemporary newspaper texts. In the case of vocabulary, the texts contain archaic words from the 19^{th} century that are more or less understandable to a contemporary reader. For example, the word *an* which can be used in the sense of "which" or "when" based on the context (in Czech *který*, *když*, respectively), the word *údové* which means "members" (contemporary Czech: *členové*) and the word *vůkol* which means "around" (*okolo*).

In the case of spelling, we can find differences that would be considered a spelling mistake today (e.g. *výtězně*, contemporary spelling: *vítězně*, "triumphantly"; *ouklady*, contemporary: *úklady*, "machinations"; *věčší*, contemporary *větší*, "bigger").

A relatively large group of differences between Czech used in our data and contemporary Czech are differences in word forms. Archaic word forms of the verb "to be" such as *býti* (infinitive), *jest* (3rd person singular) or *jsouť* (at the beginning of a sentence in the form of a particle) are used regularly in the texts. Similarly, the verb form for infinitive ending with *ti* is the only one that appears in the texts (e.g. *chrániti* - "to protect", *docíliti* - "to achieve" or *pěstovati* - "to cultivate") in comparison to contemporary regular ending for verbs *t* (contemporary: *chránit*, *docílit* or *pěstovat*, respectively). Moreover, transgressive verb forms which are now also considered archaic occur more often in our texts than in contemporary ones. For instance, *sestoupivše do spolku* - "joined the association", *vynášeje* - "bringing out" or *jdouc* - "going".

We can find different forms not only for verbs but also for nouns: *občanstvo* - "citizens" (in comparison to *občané*), *zástupcové* - "representatives" (*zástupci*); adjectives: *žádoucno* - "desirable" (in comparison to *žádoucí*) and pronouns *všickni* -"everyone" (*všichni*) and *kteráž* - "who" (*která*).

Finally, if we compare the word order in our historical newspaper texts with the contemporary ones, we can say that the contemporary word order regularly uses adjectives as premodifiers e. g.: *německá říše* - "German Empire". However, in our texts, we can find these adjectives more often as postmodifiers, e.g.: *říše německá*. Also, the position of the predicate (*byly vyschlé*) is usually at the end of a sentence in our texts (*dodržujícím parnem jako troud vyschlé byly* - "they were dried up to cinder by steady steam"), but nowadays, we would rather write *dodržujícím parnem byly vyschlé jako troud*.

In addition to the examples above, we can also find archaic abbreviations in our texts. They include different time expressions e.g. *14. t. m.* in full words *14. tohoto měsíce* which means "14th of this month", similar *t. r.* in full words *tohoto roku* - "this year". A specific abbreviation is also *pp.* which is the plural form of the abbreviation *p.* ("Mr."), *c. k.* which means *císařsko-královský* ("imperial-royal") which was used in titles of organizations in the Austrian part of Austria-Hungary in the second half of the 19^{th} century and name of currency *zl. r. č.* in full words *zlatých rakouského čísla* - "gold of Austrian number". These abbreviations are usually part of NEs and could be problematic to be automatically detected.

The texts also contain words and sentences which are not in Czech. There are some quotations in German and Latin. However, the amount of these words (sentences) is negligible.

3.2. Defining Entities

The named entity types (NEs) that we want to be able to identify are inspired by the Czech Named Entity Corpus (CNEC) (Kravalová and Žabokrtský, 2009) and their NE classification system, and adapted to the requirements of our project. We also take into consideration the special characteristics of historical texts.

CNEC presented a two-level annotation system which has 10 basic types of entities in the first level (*Numbers in addresses*, *Bibliographic items*, *Geographical names*, *Institutions*, *Media names*, *Specific number usages*, *Artifact names*, *Personal names*, *Quantitative expressions*, *Time expressions*).

On the other hand, previous research on NER for historical texts usually used only three types: *person*, *location* and *organization* (e.g., Rodriguez et al. (2012), Mac Kim and Cassidy (2015) and Neudecker (2016)). Furthermore, some of the entity types according to CNEC are not common in our texts or irrelevant for future usage of the named entities in our project (*Media names*, *Specific number usage*, *Quantitative expressions*).

Based on the discussion above, we defined five basic types of NEs, and one additional tag for ambiguity during annotation, as detailed in Table 1. This table shows the named entities, the corresponding tags and a description of what they include. We believe

Named entity	Tag	Description
Personal names	**p**	first names, surnames, artistic names, (academic) titles, (royal) family names
Institutions	**i**	names of institutions, organizations, clubs, companies, names of historical collectives (e. g. religious orders)
Geographical names	**g**	names of continents, states, territorial-administrative units, streets and public places, natural monuments including local names
Time expressions	**t**	dates, days, hours, months, years, centuries, names of epochs, holidays and important days, historic events
Artifact names / Objects	**o**	names of documents, artworks, products, books, newspapers, buildings, currency
Ambiguous	**a**	used in case the annotator is not sure which of the types above is correct

Table 1: Named entities, the corresponding tags and the description

this level of NE type classification to be sufficient for the project. It can however be extended with further level/s in the future, if necessary.

3.3. Data Preprocessing

We used optical character recognition, more specifically Tesseract 4.0[4], to transform the scanned documents to a digitized plain text format. In order to minimize the efforts of manual error correction, we omit a few pages with low scan quality for this task.

The next step was to automatically delete non-existing characters and other symbols which are never used in the original texts, because these were OCR errors: €, $, £, #, *, +, =, ©, <, >, |, [,], >> and <<. We also joined words that were separated at the end of a line and we automatically tokenized the text. Also, each sentence was separated by an empty line. This resulted in a preprocessed data set prepared for annotations, composed of 73,647 tokens.

3.4. Format and Annotation

We used an adapted CoNLL format[5] as a data structure, because it is easy to read and potential modification would also be very simple, by adding additional columns or features to the original structure. In this format, every token is placed in a separate line and additional information (lemma, morphology, language, etc.) is added in columns to the right of the token, each column separated by a space. In our case, each line contains four columns. The first one is the token, whereas the second one is reserved for lemma (represented by an underscore symbol if not specified). The third column contains information about the language. Most tokens are Czech ones ("CZ"), but we can also find some tokens in German ("DE"), French ("FR") or Latin ("LA"). The last column is used to describe the named entity type. Table 2 shows an annotation example from the corpus in this format, with an English translation of the tokens added to the very right.

[4]https://github.com/tesseract-ocr/tesseract

[5]https://www.clips.uantwerpen.be/conll2003/ner/

Token	Lemma	Lang	NE	English
zvoleni	_	CZ	O	elected
důvěrníci	_	CZ	O	confidants
k	_	CZ	O	for
volbě	_	CZ	O	the election
poslance	_	CZ	O	of deputy
pánové	_	CZ	O	gentlemen
Josef	_	CZ	B-p	Josef
Ludvík	_	CZ	I-p	Ludvík
a	_	CZ	O	and
Václav	_	CZ	B-p	Václav
Ebenstreit	_	CZ	I-p	Ebenstreit

Table 2: Example of named entity annotation in the final corpus, using the adapted CoNLL format, and with an English translation of the tokens.

Following the CoNLL format, we also used "IOB" notations to indicate the first word in a multiword entity (tag "B" for "beginning"), and inside words for all other NE units (tag "I" as "internal"). All tokens that are not a named entity are tagged as "O" - "outside" (Ramshaw and Marcus, 1995).

In our example in Table 2, we can see two personal names *Josef Ludvík* and *Václav Ebenstreit* with the tags in the fourth column *B-p* and *I-p*.

The data were manually annotated by two trained annotators. We counted inter-annotation agreement, for which we obtained 86% Cohen's kappa. Moreover, we got an accuracy of agreement in 97,3% of the tokens, taking into consideration both whether a token is to be identified as a named entity and the correct NE type. The few differences in the annotation were resolved as an agreement of both annotators (after their discussion). The annotators have also discussed ambiguous labels ("a" tag) to decide the correct one. One concrete label was selected in all cases and therefore "a" tag is not available in Table 3.

Our final corpus includes 4,017 NEs, including both one token entities and NEs consisting of more than one token. Thus 8,251 tokens from our corpus are tagged with one of our 6 NE types. An overview of the number

NE type	Tag	Token #	NE #
Personal names	p	2619	1292
Institutions	i	926	286
Geographical names	g	1335	1104
Time expressions	t	1516	506
Artifact names / Objects	o	1855	829
All		8251	4017

Table 3: Numbers of tagged tokens and NEs of each NE type in our corpus

of NEs and tagged tokens for each NE type is presented in Table 3.

The annotated corpus (73,647 tokens) was divided into three sub-sets: 80% for training, 10% for development and 10% for evaluation. Statistical information about the corpus is depicted in Table 4.

	Token #	Entity #
Training	58,984	3,188
Development	7,358	408
Testing	7,305	421
Total	73,647	4,017

Table 4: Corpus statistical information

4. Recurrent Neural Network Model

We created a simplified recurrent neural network inspired by Chiu and Nichols (2016), who obtained state-of-the-art results for the English CoNLL-2003 data set. Based on their experiments, we remove character level CNN, additional word level and lexicon based features which increase significantly the complexity of the network with only minimal impact on the resulting recognition accuracy.

Our model uses an embedding layer to map words into real numbers. All words in the vocabulary are mapped into n-dimensional embedding vectors. Our vocabulary contains 17,897 lower-cased words and the embedding vector has 300 dimensions. These can be randomly initialized and fine-tuned during the training process, set statically by some pre-trained semantic model or initialized with that model and fine-tuned during the training process. Therefore, we used the embedding vectors for contemporary Czech provided by the fastText library[6]. These vector representations are fed into a recurrent neural network (RNN), LSTM or bidirectional LSTM (BiLSTM) in our case. Finally, we use a log-softmax activation layer to get a named entity tag score. The model architecture is depicted in Figure 1.

The training of the model is realized on a per-sentence level and we used zero vectors as initial states of the RNN. For training, we applied the Nadam optimizer (Nesterov-accelerated adaptive moment estimation) (Ruder, 2016), which performed slightly better in the preliminary experiments than both the mini-batch SGD (stochastic gradient descent) and the Adam

[6]https://fasttext.cc/

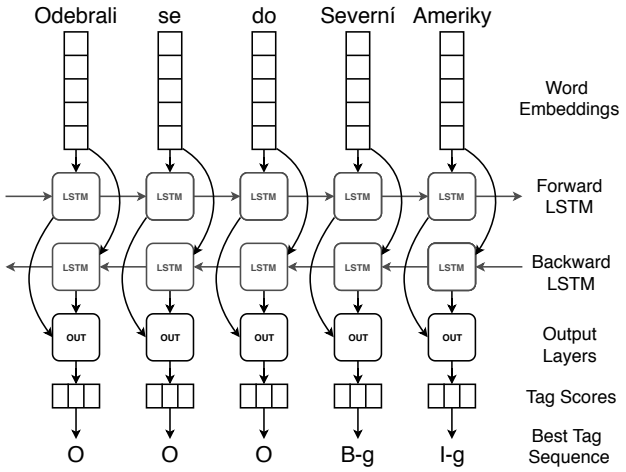

Figure 1: Model architecture

(adaptive moment estimation) optimizers. We also implemented one dropout on the input to each LSTM block, to avoid overfitting.

5. Evaluation

We evaluate our system using the standard metrics *precision*, *recall* and *F1 score*. We consider a named entity recognition as correct only in cases where both the span of the named entity and the type are correct. We also provide a qualitative analysis for some linguistic phenomena to explore more deeply the problems involved in automatic recognition of historical NEs.

5.1. Default Parameter Settings

In the first experiment, we evaluate the performance of the proposed model with similar hyper-parameter settings as suggested by Chiu and Nichols, except that the *learning rate* is set to 0.001 and that the *optimizer* used is Nadam. The aim of this experiment was to show the robustness of the model against the changes in the data set (different language, different time period).

We evaluate and compare the performance of LSTM and BiLSTM models in three different scenarios: 1) with randomly initialized embeddings; 2) with static fastText embedding vectors and 3) with dynamic fastText embeddings fine-tuned during the training of the network. The results are shown in Table 5.

As seen from the table, the BiLSTM model slightly outperforms the simpler LSTM model. Moreover, using static fastText embeddings significantly increased

Model	Precision	Recall	F1
LSTM + rand	0,556	0,518	0,536
LSTM + emb	0.747	0.609	0.671
LSTM + tune emb	0.573	0.551	0.562
BiLSTM + rand	0.611	0.492	0.545
BiLSTM + emb	0.726	0.642	**0.682**
BiLSTM + tune emb	0.648	0.589	0.617

Table 5: Precision, Recall and F1 score for LSTM and BiLSTM models with randomly initialized embeddings (+ rand), static fastText embeddings (+ emb) and dynamic fastText pre-trained embeddings fine-tuned during network training (+tune emb)

the results. However, further fine-tuning of these embedding vectors do not have a positive impact on the recognition scores.

5.2. Hyper-parameter Optimization

The second experiment aims at finding the best possible recognition score for the Czech historical corpus. We perform hyper-parameter optimization using the development set and varying the hyper-parameter values in the defined intervals. In order to support the significance of this experiment, we run it five times with a given hyper-parameter value and calculated an F1 average of each run.

We use the best performing model from the previous experiment, i.e., the BiLSTM model using static fastText pre-trained embeddings. The results are shown in Table 6 and Figure 2.

Hyper-parameter	Our corpus		CoNLL
	Range	Final	
LSTM state #	[100; 500]	250	275
LSTM layer #	[1; 3]	1	1
Learning rate	[0.001; 0.01]	0.004	0.0105
Epochs	see Fig. 2	80	80
Dropout	[0.25; 0.85]	0.65	0.68

Table 6: Overview of hyper-parameter optimization in comparison to the settings proposed by Chiu and Nichols on English CoNLL-2003 data.

The first column shows the hyper-parameters to optimize, whereas the second column describes the intervals, and the third column depicts the parameter value to obtain the highest recognition score on the development set. The last column shows the original hyper-parameters values used by Chiu and Nichols on the English CoNLL-2003 data. The results show that our best hyper-parameters slightly differ from those proposed by Chiu and Nichols.

5.3. Final Results

Our final experiment aims at presenting the best achieved recognition scores with the hyper-parameters previously fine-tuned. We also depict the results of our BiLSTM models on English CoNLL-2003 data to compare the performance of our model on another data set

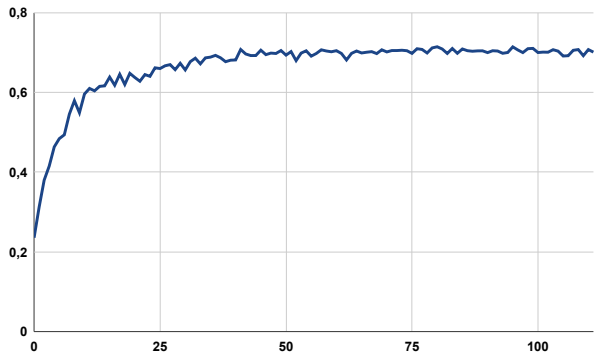

Figure 2: Learning curve of BiLSTM model with static pre-trained fastText embeddings; X axis shows the number of training epochs, while Y axis represents the recognition score

(different language, different time period) with previous work.

As previously, we evaluate our model in three different embedding settings. In the case of the Czech data we use the fastText word embeddings, whereas for English, we employ 50-dimensional Glove[7] word vectors. Based on the experiments, we use also for English the different learning rate of 0.001. The results of this experiment are detailed in Table 7.

The best F1 score obtained for the Czech historical corpus was 0.73 using the BiLSTM model with static pre-trained fastText word embeddings. This model with a similar embeddings setup also obtained the best results on the English CoNLL-2003 data set, with an F1 score of 0.879.

If we compare the results of our BiLSTM model implementation on the CoNLL-2003 data sets with randomly initialized embeddings (F1 score 0.764) with the results of Chiu and Nichols (F1 score 0.763) with a similar embeddings setup we see that our model reached comparable results.

5.4. Qualitative Analysis

For qualitative analysis, we compared the tagged output text of the entire evaluation data set to the same text in its manually annotated version. The goal of the analysis was to observe in which cases the model automatically detects the NEs with correct tags and under what circumstances it does not. We also tried to describe the linguistic phenomena of historical Czech which cause problems for our BiLSTM model using static pre-trained fastText word embeddings for contemporary Czech.

The analysis shows that the model gives satisfying output for the following NE types:

> *time expressions*, especially in format *12. července 1872* ("12th July 1872")
>
> *geographical names*, especially names of cities such as *Domažlice*, *v Trhanově* ("in Trhanov"), *z*

[7] https://nlp.stanford.edu/projects/glove/

Model	Our corpus			CoNLL-2003		
	Precision	Recall	F1	Precision	Recall	F1
BiLSTM + rand	0.630	0.518	0.568	0.825	0.712	0.764
BiLSTM + emb	0.764	0.698	**0.730**	0.879	0.879	**0.879**
BiLSTM + tune emb	0.654	0.599	0.625	0.848	0.876	0.862

Table 7: Final results of the selected neural models on the Czech historical NER corpus and on CoNLL-2003 data sets with randomly initialized embeddings (+ rand), static embeddings (+ emb) and dynamic pre-trained embeddings fine-tuned during training; fastText is used for our corpus, Glove for English.

Petrohradu ("from Petrohrad") or *do Solnohradu* ("to Solnohrad"). We included prepositions in the examples to show that they help the model to detect these NEs correctly.

personal names, especially quite common Czech names or names common in input data sets, e.g. *Antonín*, *Jiří Prunař* or *Dr. Ant. Steidl*,

artifact names, especially abbreviated names of currencies, e. g. *zl.* (*zlatý*), *kr.* (*krejcar*) and *zl. r. č.* (*zlatý rakouského čísla*, "gold of Austrian number").

institutions including multiword NEs, e.g.: *Jenerální zastupitelství rakouského ústředního stavitelského spolku ve Vídni* ("General Assembly of the Austrian Central Building Society in Vienna") or *knihkupectví Jiř. Prunara* ("Bookstore of Jiř. Prunar").

In case of *institutions*, tokens such as *ústav* ("institute"), *obecní* ("municipal") or *pivovar* ("brewery") are tagged as NE type *institution* because they are often part of institution names. These problems are caused by the fact that the names of institutions are usually multiword expressions, therefore harder to detect correctly. Our model also tagged as *institution* completely wrong word sequences as *Finanční ministr předložil zákon* ("The Finance Minister has submitted a law"), where words *Finanční* and *ministr* could signal the name of an institution.

Our model also had problems to correctly detect some of the other *geographical names*, especially, if a previous token of the NE is an uncommon preposition in Czech. For example, in the word sequence *v Čechách, na Moravě a v Slezsku* ("in Bohemia, in Moravia, in Silesia"), the tokens *Čechách* and *Slezsku* were tagged correctly but the token *Moravě* was not tagged at all. Similarly, the word sequence *na Brodě* ("in Brod") that starts with preposition *na* was tagged incorrectly. Moreover, if a part of a multiword expression is common in other names, the model tags the part correctly. For example, the token pair *Skleněná Huť*, where the second one is more common and is tagged correctly, whereas the first token is not.

Moreover, the model did not detect some *personal names* and *artifact names* at all. These names are not common in our training data set, e.g. *Christian Kotz*, *Wertzlera* or *Jelínek* and *Osvěta* or *Dle přírody* (names of newspapers and a book), respectively.

In the case of *time expressions*, we found that the model has problems to correctly detect expressions including names of months in word forms which are not so frequent in the Czech language, e.g. *1. srpnem* (more common word form *srpna*, "25th August") or *června 1872* (more common word form *červen*, "June") were not tagged.

Naturally, the model tagged some tokens which are not NEs, e.g. *továrníka* ("factory owner") or *hlas* ("voice"), some NEs were tagged with the wrong NE type, e.g. *Trojan* (personal name) or *sv. Kříže* ("Holy Cross") or they were not tagged at all, e.g. *Radnice* (name of town).

To sum up, based on the analysis, we can say that the issues described above are not caused only by the fact that the input data were historical texts. These issues could be partly solved by larger input data sets therefore our results are promising for future research.

6. Conclusions and Future Work

In this paper, we introduced the *Czech historical named entity corpus v 1.0*, a novel collection of annotated texts for historical Czech NER which is freely available for research purposes. We also show several experiments using LSTM and BiLSTM methods.

The corpus is composed of Czech text newspapers from the second half of 19^{th} century. We specified the basic NE-types: *Personal names*, *Institutions*, *Geographical names*, *Time expressions* and *Artifact names / Objects* and created an annotation manual with concrete examples for our final corpus. This manual is provided within the corpus.

We conducted several experiments on this data set using LSTM and BiLSTM recurrent networks. We also experimented with three different word embedding approaches. We have shown that a BiLSTM model using static pre-trained fastText word embeddings achieved the best performance with an F1 score of 0.73. We can conclude that the pre-trained embeddings improve the results even when using contemporary Czech for training. We also evaluated the model on the English CoNLL–2003 data set, where our results are comparable with previous work.

Moreover, we provided a qualitative analysis of the observed linguistic phenomena, where we identified that the BiLSTM model has difficulties to detect words, word forms and specific sequences of tokens that are uncommon in our training data set.

In our future research, the first task consists in the lemmatization and morphological analysis of the cor-

pus. This information will be provided in the further version.

Then, we will propose and evaluate more complex neural net topologies including, for instance, attentive or gated recurrent networks.

The last task is the experimentation with more sophisticated word embeddings (see Devlin et al. (2019) or Peters et al. (2018)) and using historical data for their training.

7. Acknowledgements

This work has been partly supported by Cross-border Cooperation Program Czech Republic - Free State of Bavaria ETS Objective 2014-2020 (project no. 211) and by Grant No. SGS-2019-018 Processing of heterogeneous data and its specialized applications.

8. Bibliographical References

Chiu, J. and Nichols, E. (2016). Named Entity Recognition with Bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 999888:2493–2537, November.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). 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)*, pages 4171–4186.

Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 363–370, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Grover, C., Givon, S., Tobin, R., and Ball, J. (2008). Named entity recognition for digitised historical texts. In *LREC 2008*.

Huang, Z., Xu, W., and Yu, K. (2018). Bidirectional LSTM-CRF models for sequence tagging. In *32nd Pacific Asia Conference on Language, Information and Computation*. Association for Computational Linguistics.

Král, P. (2011). Features for named entity recognition in czech language. In *In proceedings international conference on knowledge engineering and ontology development. Setúbal: SciTePress*, pages 437–441.

Kravalová, J. and Žabokrtský, Z. (2009). Czech Named Entity Corpus and SVM-based recognizer. In *Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration (NEWS 2009)*, pages 194–201. Association for Computational Linguistics.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (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.

Mac Kim, S. and Cassidy, S. (2015). Finding names in trove: named entity recognition for Australian historical newspapers. In Ben Hachey et al., editors, *Australasian Language Technology Association Workshop 2015*, volume 13, pages 57–65. Australasian Language Technology Association.

Neudecker, C. (2016). An open corpus for named entity recognition in historic newspapers. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Packer, T. L., Lutes, J. F., Stewart, A. P., Embley, D. W., Ringger, E. K., Seppi, K. D., and Jensen, L. S. (2010). Extracting person names from diverse and noisy OCR text. In *AND*, pages 19–26. ACM.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237.

Ramshaw, L. and Marcus, M. (1995). Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

Rodriguez, K. J., Bryant, M., Blanke, T., and Luszczynska, M. (2012). Comparison of named entity recognition tools for raw OCR text. In *Proceedings of KONVENS 2012 (LThist 2012 workshop)*.

Ruder, S. (2016). An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*.

Straková, J., Straka, M., and Hajič, J. (2014). Open-source tools for morphology, lemmatization, POS tagging and named entity recognition.

Straková, J., Straka, M., and Hajič, J. (2013). A new state-of-the-art Czech named entity recognizer. In Ivan Habernal et al., editors, *TSD*, volume 8082 of *Lecture Notes in Computer Science*, pages 68–75. Springer.

Ševčíková, M., Žabokrtský, Z., and Kruza, O. (2007a). Named entities in Czech: Annotating data and developing NE tagger. volume Lecture Notes in Artificial Intelligence. Proceedings of the 10th International Conference Text, Speech and Dialogue (TSD 2007), pages 188–195.

Ševčíková, M., Žabokrtský, Z., and Kruza, O. (2007b). Zpracování pojmenovaných entit v českých textech. ÚFAL MFF UK.