Alleviating Digitization Errors in Named Entity Recognition for Historical Documents

Emanuela Boros[1], **Ahmed Hamdi**[1], **Elvys Linhares Pontes**[1], **Luis Adrián Cabrera-Diego**[1], **Jose G. Moreno**[1,2], **Nicolas Sidere**[1], and **Antoine Doucet**[1]

[1] University of La Rochelle, L3i, F-17000, La Rochelle, France
{emanuela.boros,ahmed.hamdi,elvys.linhares_pontes,luis.cabrera_diego}@univ-lr.fr
{nicolas.sidere,antoine.doucet}@univ-lr.fr

[2] University of Toulouse, IRIT, UMR 5505 CNRS, F-31000, Toulouse, France
jose.moreno@irit.fr

Abstract

This paper tackles the task of named entity recognition (NER) applied to digitized historical texts obtained from processing digital images of newspapers using optical character recognition (OCR) techniques. We argue that the main challenge for this task is that the OCR process leads to misspellings and linguistic errors in the output text. Moreover, historical variations can be present in aged documents, which can impact the performance of the NER process. We conduct a comparative evaluation on two historical datasets in German and French against previous state-of-the-art models, and we propose a model based on a hierarchical stack of Transformers to approach the NER task for historical data. Our findings show that the proposed model clearly improves the results on both historical datasets, and does not degrade the results for modern datasets.

1 Introduction

With the emergence of large scale archives of digitized contents, the need for efficient preservation and accessibility of historical documents through appropriate technologies increased exponentially. At the same time, there is a growing interest in extracting relevant information from historical sources. In this paper, we address the named entity recognition (NER) task which aims at identifying real-world entities, such as names of people, organizations, and locations within historical documents.

Since most of the state-of-the-art research focuses on NER for modern available datasets, the performance of the NER systems grew at a fast pace, enabled by the representational capacity of neural networks and off-the-shelf pre-trained word embeddings (Ma and Hovy, 2016; Lample et al., 2016; Yadav and Bethard, 2018). More recently, NER models based on contextual word and subword representations provided by ELMo (Peters et al., 2018), Flair (Akbik et al., 2018), or BERT (Devlin et al., 2019), achieved impressive improvements. The Transformer-based (Vaswani et al., 2017) architectures for NER became popular since the release of the BERT (Bidirectional Encoder Representations from Transformers) model.

However, while most NER systems have been developed to generally address contemporary data, NER systems for processing historical documents are less common. To extract entities from historical documents, NER tools face additional challenges. As the majority of these documents are hardcover, they are scanned and processed by an OCR to transcribe the text. However, an OCR tool can occasionally misrecognize letters and improperly identify its textual content. This can be due to the level of degradation of the actual document being scanned, to the digitization artifacts and also to the quality of the OCR tool. This leads to digitization errors in the transcribed text, such as misspelled locations or person names.

Languages evolve through time and certain words can have a different meaning depending on the period of time analyzed (Hamilton et al., 2016). The spelling of words can also change due to new orthographic conventions or cultural tendencies (Scheible et al., 2011). This high level of spelling differences can be incompatible with modern orthography and the produced noise can severely affect modern NLP systems (Lopresti, 2009).

To address these challenges of NER on historical documents, we propose a robust NER model based on a stack of Transformers that includes fine-tuned BERT encoders. We study the impact of such a model, and we conclude that this type of model is suited for the extraction of entities from historical documents.

Proceedings of the 24th Conference on Computational Natural Language Learning, pages 431–441
Online, November 19-20, 2020. ©2020 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17

The remainder of this paper is organized as follows. In Section 2, we present and discuss a selection of works concerning NER in modern and historical documents. Then, in Section 3, the datasets explored in this work are presented. The proposed model is detailed in Section 4. The experiments are described in Section 5. We present and discuss the obtained results in Section 6. Finally, Section 7 concludes this paper and hints at future work.

2 Related Work

NER for modern documents The first end-to-end systems for sequence labeling tasks are based on pre-trained word and character embeddings encoded either by a bidirectional Long Short Term Memory (BiLSTM) network or a Convolutional Neural Network (CNN) (Collobert et al., 2011; Lample et al., 2016; Ma and Hovy, 2016; Aguilar et al., 2017; Chiu and Nichols, 2016), along with a Conditional Random Fields (CRF) decoder. One shortcoming of this type of model is that they were based on a single context-independent representation for each word. This problem has been further attenuated by methods based on language model pre-training that produced context-dependent word representations. These recent large-scale language models methods such as BERT (Devlin et al., 2019) and ELMo (Peters et al., 2018) further enhanced the performance of NER, yielding state-of-the-art performances (Peters et al., 2017, 2018; Baevski et al., 2019).

NER for historical documents Historical documents pose multiple challenges that either depend on the quality of digitization or the historical variations of a language. Studies on how the NER models can be impacted by the digitization process (Miller et al., 2000; Rodriquez et al., 2012; Hamdi et al., 2019; van Strien et al., 2020) have clearly shown that the performance scores of a NER model can significantly decrease when applied on historical documents.

The increased interest in contributing to historical language resources is driven forward by the creation of new gold standards for historical document processing. For example, Hubková (2019) created and annotated a corpus using scanned Czech historical newspapers, and Ahmed et al. (2019) proposed a German gold standard for NER in historical biodiversity literature.

A recent competition organized by the *Identifying Historical People, Places, and other Entities* (HIPE) lab at CLEF 2020[1], not only that it created a gold standard for German and French historical texts, but also encouraged researchers to participate in two sub-tasks, named entity recognition and classification and entity linking.

Considering the high level of spelling differences between modern and historical documents, variance (inconsistency), and uncertainty (digitization errors) found in historical documents, the recent methods assess these shortcomings differently.

Erdmann et al. (2016) presented a CRF-based model with handcrafted features for Latin historical texts and motivated the choice of Part-of-Speech (POS) tagger by the fact that this NLP tool leverages the highly informative morphological complexity of Latin. The BiLSTM-based model proposed by Hubková (2019) applied a character-based CNN to encode the different spellings of words.

Similar to the latter approach, we also consider that the NER model itself can help in alleviating the historical documents issues, without the use of language-specific engineered features. Differently, we introduce the NER for historical documents to the language model methods based on the Transformer architecture (Vaswani et al., 2017) and BERT (Devlin et al., 2019) methods, that, to our knowledge, have not been approached in previous research, with regard to processing historical documents.

With new needs and resources in the context of historical NER processing, we evaluate our proposed model on the dataset proposed by the HIPE competition, and we also propose a new gold standard for German and French, to assess our assumptions.

3 Datasets

We conduct experiments on two datasets that comprise digitized historical newspapers, HIPE and NewsEye datasets in French and German. Additionally, we study how the proposed methods behave in the case of contemporary data, by experimenting on the English CoNLL 2003 dataset (Tjong Kim Sang and De Meulder, 2003).

The HIPE dataset was created by the CLEF 2020 Evaluation Lab HIPE challenge (Ehrmann et al., 2020a). It is composed of articles from several Swiss, Luxembourgish, and American historical newspapers from 1790 to 2010 (Ehrmann et al.,

[1] impresso.github.io/CLEF-HIPE-2020/

2020b). More concisely, the German articles were collected from 1790 to 1940, and the French articles, from 1790 to 2010. The corpus was manually annotated by natives following the annotation guidelines derived from the Quaero annotation guide[2].

We also present the NEWSEYE dataset, composed of historical newspapers in French (1814-1944) and German (1845-1945). The documents were collected through the national libraries of France[3] (BnF) and Austria[4] (ONB), respectively. This dataset was annotated following guidelines derived from the Quaero annotation guide[5]. The annotation process was made by native speakers for each language using the Transkribus tool[6]. In order to compute the inter-annotator agreement (IAA), we used the Kappa coefficient introduced by Cohen (1960). Several pages from each corpus (German and French) have been annotated twice by two groups of annotators. Satisfactory IAA scores were reached for the two corpora (0.90 for French and 0.91 for German). The NewsEye corpus is split into 80% for training and 20% for both validation and testing.

The CoNLL 2003 dataset consists of newswire from the Reuters RCV1 corpus and it includes standard train, development, and test sets.

Table 1 presents the statistics regarding the number and type of entities in the aforementioned datasets. The statistics are divided according to the training, development, and test sets.

		FR			DE		
	Type	train	dev	test	train	dev	test
HIPE	LOC	3,067	664	854	1,747	771	595
	ORG	833	172	130	358	158	130
	PERS	2,513	428	502	1,170	677	311
	PROD	198	53	61	112	48	62
	TIME	273	73	53	118	69	49
NEWSEYE	LOC	4,878	522	698	4,024	525	894
	ORG	1,602	142	229	3,171	307	252
	PERS	5,023	853	788	2,346	424	461
	PROD	185	57	23	43	12	16

		EN		
	Type	train	dev	test
CoNLL-03	LOC	7,140	1,837	1,668
	ORG	6,321	1,341	1,661
	PERS	6,600	1,842	1,617
	MISC	3,438	922	702

Table 1: Overview of the HIPE, NEWSEYE, and CoNLL 2003 datasets statistics. LOC = Location, ORG = Organization, PERS = Person, PROD = Product, TIME = Time and MISC = Miscellaneous.

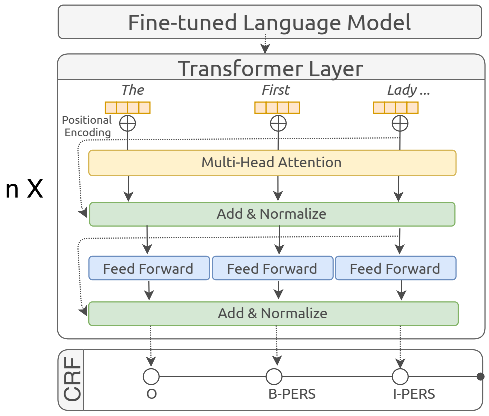

Figure 1: Main architecture of the BERT+n×Transf.

4 Model

We based our NER model on the pre-trained model BERT proposed by Devlin et al. (2019). Although original recommendations suggest that unsupervised pre-training of BERT encoders are expected to be sufficiently powerful on modern datasets, we consider that adding extra Transformer layers could contribute to the alleviation of word errors or misspellings.

First, we use a pre-trained BERT model, and second, we stack n Transformer blocks on top, finalized with a CRF prediction layer. We refer to this model as BERT+n×Transf where n is a hyperparameter referring to the number of Transformer layers. The global architecture of our model is depicted in Figure 1. We used Transformer blocks with parameters that we chose empirically similar to the configuration of the blocks in the fine-tuned model[7].

The reasons for using BERT models are that they can easily be fine-tuned for a wide range of tasks, but also that they produce high-performing systems (Devlin et al., 2019; Conneau and Lample, 2019; Radford et al., 2018). Nonetheless, despite the major impact of BERT in the NLP community, re-

[2] Quaero guidelines

[3] https://www.bnf.fr

[4] https://www.onb.ac.at/

[5] The main difference is that several named entities subtypes were ignored. In addition, the TIME type was not included in the annotation of the NEWSEYE dataset.

[6] https://transkribus.eu/Transkribus/

[7] Note that they can vary as multiple BERT-based models are available for different languages.

searchers question the ability of this model to deal with noisy text (Sun et al., 2020) unless complementary techniques are used (Muller et al., 2019; Pruthi et al., 2019).

More specifically, the built-in tokenizer of BERT first performs simple white-space tokenization, then applies a Byte Pair Encoding (BPE) based tokenization, WordPiece (Wu et al., 2016). For example, word can be split into character n-grams (e.g. compatibility → 'com', '##pa', '##ti', '##bility'), where ## is a special symbol for representing the presence of a sub-word that was recognized.

Between the types of OCR errors that can be encountered in historical documents, the character insertion modification has the minimum influence (Sun et al., 2020), because the tokenization at the sub-word level of BERT would not change much in some cases, such as 'practically' → 'practicaally'. Meanwhile, the substitution and deletion errors can hurt the performance of the tokenizer the most due to the generation of uncommon samples, such as 'professionalism' → 'pr9fessi9nalism' that is tokenized as 'pr', '##9', '##fes', '##si', '##9', '##nal', '##sm'. BERT has been demonstrated to have a sensitivity to its sub-word segmentation when it comes to such words, as the meaning of the sub-words can diminish the initial meaning of the correctly spelled word (Sun et al., 2020). Thus, these new noisy tokens could influence the performance of BERT-based models[8].

On top of BERT, we add a stack of Transformer blocks (encoders). A Transformer block (encoder), as proposed in (Vaswani et al., 2017), is a deep learning architecture based on multi-head attention mechanisms with sinusoidal position embeddings. It is composed of a stack of identical layers. Each layer has two sub-layers. The first is a multi-head self-attention mechanism, and the second is a simple, position-wise fully connected feed-forward network. A residual connection is around each of the two sub-layers, followed by layer normalization. All sub-layers in the model, as well as the embedding layers, produce outputs of dimension 512. In our implementation, we used learned absolute positional embeddings (Gehring et al., 2017) instead, as it is a common practice[9]. Vaswani et al. (2017) found that the two versions produced nearly identical results.

We assume that the additional Transformer layers can alleviate the sensitivity of the built-in tokenizer of BERT towards OOV, OCR errors, or misspellings, and contribute to the learning or finding the proper informative words around entities.

5 Experiments

5.1 Baseline

We chose as a baseline the model proposed by Ma and Hovy (2016), an end-to-end model combining a BiLSTM and a CNN character encoding, in order to take advantage of the word and character features. The character-level features are known to capture morphological and shape information (Kanaris et al., 2007; Santos and Zadrozny, 2014; dos Santos and Guimarães, 2015) that can also offer the possibility of obtaining a representation for misspelled, custom, or abnormal words. For the baseline, we used the FastText[10] pre-trained word embedding models (Grave et al., 2018)[11].

Additionally, we analyze the aid that can be brought by an available larger dataset by training the baseline model in two stages in a transfer learning setting, similar to the setting in which the BERT encoder is used in our model:

1. *pre-training*, where the network is trained on a larger-scale available contemporary dataset
2. *fine-tuning*, where the pre-trained network is further trained on the historical datasets

The modern datasets are the following:

- For French, we use the fr-WikiNER[12] dataset that is extracted from Wikipedia articles. It contains about 500k tokens from which around 31k are named entities.
- For German, we use the de-GermEval[13] dataset generated from German Wikipedia and News Corpora as a collection of citations. The dataset covers over 31k sentences corresponding to over 590k tokens from which around 33k are named entities.

[8] To increase the chances for misspelled, non-canonical, or new words to be recognized, we enrich the vocabulary of the tokenizer with these tokens, while allowing not only the BERT encoder but also the added Transformer layers to learn them from scratch.

[9] https://huggingface.co/

[10] https://fasttext.cc/docs/en/crawl-vectors.html

[11] For a more detailed description of the model and of the hyperparameters can be found in Ma and Hovy (2016).

[12] https://figshare.com/articles/Learning_multilingual_named_entity_recognition_from_Wikipedia/5462500

[13] https://sites.google.com/site/germeval2014ner/data

5.2 Metrics

The evaluation of the NER task is done in a coarse-grained manner, with the entity (not token) as the unit of reference (Makhoul et al., 1999). We compute precision (P), recall (R), and F1 measure (F1) at micro-level, i.e. error types are considered over all documents. Two evaluation scenarios were considered: *micro-strict*, which looks for an exact boundary matching, and *micro-fuzzy*, where a prediction is correct when there is at least one token overlap (Ehrmann et al., 2020a). Further, statistical significance is measured through a two-tailed t-test, with an estimated p-value between 0.01 and 0.05.

5.3 Data Pre-processing

The HIPE dataset was initially segmented at the article-level. Since BERT is able to consume only a limited context of tokens as their input (512), we segment the articles at sentence-level. We also reconstruct the original text, including hyphenated words. The reconstructed text was passed through Freeling 4.1 (Padró and Stanilovsky, 2012) to obtain a segmentation based on sentences. We made use of the same segmentation for the baseline model. Moreover, for the BERT+n×Transf, we feed the model with batches of same sized inputs.

5.4 Hyperarameters

The hyperparameters used for both models are depicted as follows.

For the German NER, we chose as a pre-trained encoder the `bert-base-german-europeana`. This BERT model has been used in other NER tasks for processing contemporary and historical German documents (Schweter and Baiter, 2019; Riedl and Padó, 2018). It was trained using a large collection of newspapers provided by the Europeana Library.[14]

For the French NER, we rely on the large version of the pre-trained CamemBERT (Martin et al., 2020) model, i.e. (`camembert-large`). This model was trained on a large French corpus. CamemBERT proposes some differences with respect to other BERT models. For instance, it uses whole-word masking and SentencePiece tokenization (Kudo and Richardson, 2018) instead of WordPiece tokenization (Wu et al., 2016) as the original BERT.

For the English dataset CoNLL, we experimented with both `bert-base-cased` and `bert-large-cased`, pre-trained models presented in (Devlin et al., 2019).

We denote the number of layers (i.e., Transformer blocks) as L, the hidden size as H, and the number of self-attention heads as A. `bert-base-cased` has L=12, H=768, A=12, `bert-large-cased` and `camembert-large`, L=24, H=1024, A=16. In all the cases, the top Transformer blocks have L=1 for 1×Transf and L=2 for 2×Transf, H=128, A=12, chosen empirically. The BERT-based encoders are fine-tuned on the task during training.

For training, we followed the selection of parameters presented in (Devlin et al., 2019). We found that 2×10^{-5} learning rate and a mini-batch of dimension 4 for German and English, and 2 for French, provide the most stable and consistent convergence across all experiments as evaluated on the development set.

6 Results

In this section, we provide experimental results of the baseline model and the proposed method. In order to assess the ability of both models with regard to the presence of errors provided by an OCR, we present several experiments:

- In Table 2, the first two experiments are performed with the baseline model, with and without the pre-training proposed by the transfer learning method on larger contemporary datasets.

- It is necessary to analyze how sensitive the proposed model is to the number of Transformer layers, the hyper-parameter n. Therefore, we conduct two experiments for ablation study with the n value $\in \{0, 1, 2\}$. The values > 2 obtained lower performance results and had a tendency to overfit. Therefore, in the same Table 2, we present next these experiments.

- In Table 3, the results for the baseline model without any transfer learning (as it was unnecessary) are presented, along with the same ablation study for the BERT+n×Transf.

From the results in the Table 2, we can see the evidence that the BERT-based models with n×Transf achieve, for both datasets and languages, higher *micro-fuzzy* and *micro-strict* performance values than the BERT model stand-alone and the baseline

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

	HIPE						**NewsEye**					
	DE			FR			DE			FR		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
BiLSTM-CNN												
fuzzy	83.3	70.1	76.1	89.9	83.9	86.8	81.2	42.4	55.7	82.2	77.2	79.6
strict	69.4	58.4	63.4	77.7	72.5	75.0	54.8	28.6	37.6	65.5	61.4	63.4
BiLSTM-CNN (transfer learning)†												
fuzzy	81.1	75.0	77.9**	87.8	88.8	88.3	76.4	49.4	60.0**	83.6	77.8	80.6*
strict	67.4	62.2	64.7**	77.3	78.2	77.7	48.6	31.4	38.1**	66.9	62.3	64.5*
BERT												
fuzzy	83.4	88.3	85.8**	89.5	91.9	90.7*	60.1	67.0	63.4**	86.1	81.8	83.9**
strict	74.1	78.5	76.2**	81.1	83.3	82.1*	46.8	52.2	49.4**	70.1	66.6	68.3**
BERT+1×Transf												
fuzzy	85.8	87.3	86.5**	91.3	92.9	**92.1****	82.3	66.4	**73.5****	88.7	82.1	**85.3****
strict	77.2	78.6	77.9**	83.5	84.9	**84.2****	62.7	50.6	56.0**	74.4	68.9	**71.5****
BERT+2×Transf												
fuzzy	87.0	87.2	**87.1****	91.5	92.4	91.9**	83.3	64.4	72.6**	89.7	80.1	84.7 **
strict	78.6	78.7	**78.7****	83.4	84.2	83.8**	64.9	50.2	**56.6****	75.0	67.0	70.8**

Table 2: NER test results for the HIPE and NewsEye datasets in French and German. All models have as a decoder layer a CRF. †= with pre-training on larger modern datasets. All metrics are micro. Statistical significance is measured through a two-tailed t-test. * denotes a significant improvement over the BiLSTM model at p $\leq$ 0.05, ** denotes p $\leq$ 0.01.

models. All models have a statistical significance < 0.01, thus, adding $n\times$Transf can improve model generalizability for NER on historical documents.

Moreover, they generally manage to maintain a balance between recall and precision, while the baseline models vary, depending on the language. We also notice that, while in general, both models obtain a more or less precision-recall balance, there are two cases where there is a large imbalance, more specifically in the NewsEye German dataset. Comparing with the baseline models, the BERT+$n\times$Transf only achieves a 20 percentage points difference between precision and recall, while the baseline suffers from 40 points difference.

In the context of transfer learning applied for the baseline models, two performance results, for NewsEye in German, and for HIPE in French are higher due to the fine-tuning on these datasets, while the others are not degraded by the pre-training on larger contemporary datasets. This observation confirms the previous studies done on this type of model regarding their robustness to misspellings (Sun et al., 2020; Pruthi et al., 2019). We also notice that for German both datasets, the results for transfer learning from contemporary German datasets are statistically significant ($< 0.01\%$), while contemporary datasets the performance difference for both French datasets was minimal (either < 0.5 for French NewsEye or < 0.9 for French HIPE).

	CoNLL-03 EN					
	P	R	F1			
BiLSTM-CNN						
micro-fuzzy	91.0	89.7	90.4			
micro-strict	89.2	87.9	88.5			
	P	R	F1	P	R	F1
	`bert-base-cased`			`bert-large-cased`		
BERT						
micro-fuzzy	91.7	93.0	92.3	92.4	93.5	92.9
micro-strict	90.3	91.6	90.9	91.1	92.2	91.6
BERT+1×Transf						
micro-fuzzy	92.5	93.2	**92.8**	92.7	93.4	**93.1**
micro-strict	91.1	91.8	**91.4**	91.4	92.1	**91.8**
BERT+2×Transf						
micro-fuzzy	92.0	93.2	92.6	92.9	93.4	**93.1**
micro-strict	90.6	91.8	91.2	91.6	92.1	**91.8**

Table 3: NER test results for the CoNLL 2003 dataset. All models have as a decoder layer a CRF.

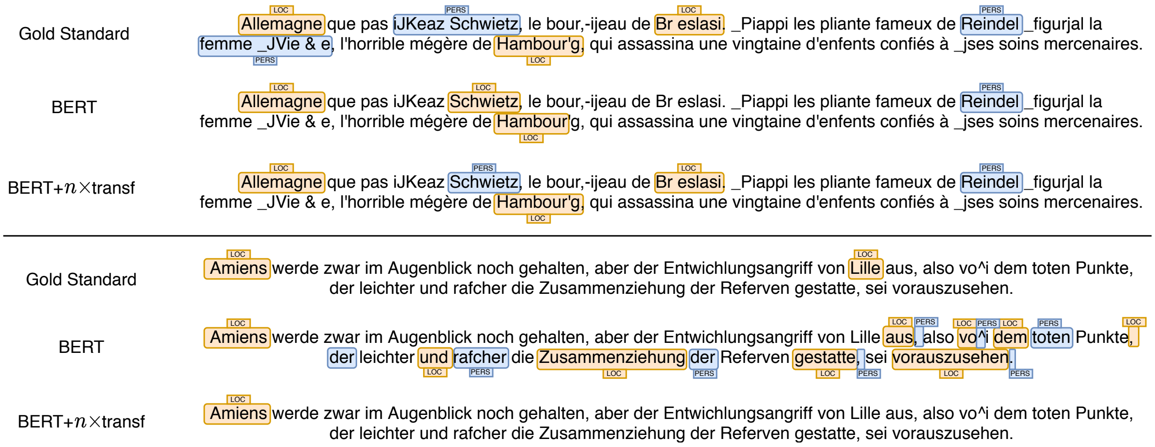
Gold Standard
LOC Allemagne que pas PERS iJKeaz Schwietz, le bour,-ijeau de LOC Br eslasi. _Piappi les pliante fameux de PERS Reindel _figurjal la PERS femme _JVie & e, l'horrible mégère de LOC Hambour'g, qui assassina une vingtaine d'enfents confiés à _jses soins mercenaires.
BERT
LOC Allemagne que pas iJKeaz LOC Schwietz, le bour,-ijeau de Br eslasi. _Piappi les pliante fameux de PERS Reindel _figurjal la femme _JVie & e, l'horrible mégère de LOC Hambour'g, qui assassina une vingtaine d'enfents confiés à _jses soins mercenaires.
BERT+n×transf
LOC Allemagne que pas iJKeaz PERS Schwietz, le bour,-ijeau de LOC Br eslasi. _Piappi les pliante fameux de PERS Reindel _figurjal la femme _JVie & e, l'horrible mégère de LOC Hambour'g, qui assassina une vingtaine d'enfents confiés à _jses soins mercenaires.
Gold Standard
LOC Amiens werde zwar im Augenblick noch gehalten, aber der Entwichlungsangriff von LOC Lille aus, also vo^i dem toten Punkte, der leichter und rafcher die Zusammenziehung der Referven gestatte, sei vorauszusehen.
BERT
LOC Amiens werde zwar im Augenblick noch gehalten, aber der Entwichlungsangriff von Lille LOC aus PERS , also LOC vo PERS ^i LOC dem PERS toten Punkte LOC , PERS der leichter LOC und PERS rafcher die LOC Zusammenziehung PERS der Referven LOC gestatte PERS , sei LOC vorauszusehen PERS .
BERT+n×transf
LOC Amiens werde zwar im Augenblick noch gehalten, aber der Entwichlungsangriff von Lille aus, also vo^i dem toten Punkte, der leichter und rafcher die Zusammenziehung der Referven gestatte, sei vorauszusehen.

Figure 2: An example of NER predictions on the HIPE dataset in French (top part) and German (bottom part).

In the context of modern data, in the Table 3, the F1 values of the stand-alone BERT model applied on the CoNLL 2003 dataset fairly correspond to the ones reported in (Devlin et al., 2019) (the authors report a F1 of 92.4% for `bert-base-cased` and 92.8% for `bert-large-cased`). While the F1 value has a very small margin difference from the (Devlin et al., 2019), the performance results for the BERT+n×Transf slightly increased for both proposed models. We assume that one reason would be that the capacity of representation of extra Transformer layers, even in a context where no misspelling errors are present, can contribute to a modest improvement. While this improvement is more visible for the BERT `bert-base-cased`+1×Transf (a difference of a half of percentage point), and 0.3 percentage points for `bert-base-cased`+2×Transf, for the `bert-large-cased` BERT+n×Transf, the values remain unchanged (with a difference of 0.2 percentage points from BERT).

6.1 Discussion

For more qualitative analysis, we examine the number of unrecognized words by the pre-trained BERT-based models that were added to the specific tokenizers (WordPiece for BERT and SentencePiece for CamemBERT). For NewsEye German, 8.84% of the total number of words in the vocabulary needed to be fully trained, while only 0.14% were unknown in the HIPE dataset. Following this observation, we notice that there is a large F1 margin between BERT+CRF and BERT+n×Transf (63.4% in comparison with 73.5% and 72.6%, respectively), a fact that could be motivated by the large percentage of unknown words.

Moreover, for German, even though the BERT encoder was pre-trained on a digitized historical dataset (`bert-base-german-europeana`), the proposed model contributed greatly to the coverage of the misspelled or abnormal words present in the NewsEye. For French, the results vary of around $1-2$ percentage F1 points between the stand-alone BERT and the BERT+n×Transf models.

Between the two datasets, only HIPE was also annotated with the Levenshtein Ratio between the gold standard entities and the transcribed ones. In Figure 3, we compare BERT and BERT+n×Transf by analyzing the number of correct predicted entities with respect to the Levenshtein distance. For the French predictions, for 56.25% of the different values of the distance, the stacked models had relatively more correct predictions. A French example of a misspelled entity that is recognized by both BERT+n×Transf but not by BERT is presented in Figure 2, in the upper part. For German, only in 18.75% of the cases, the stacked models have more correctly identified entities that are misrecognized.

We also presume that the introduction by the stacked Transformers of additional hyperparameters can increase the ability of the architecture to better model long-range contexts. Thus, we analyzed the correctly predicted German and French HIPE entities by their length. We noticed that BERT+n×Transf is better than BERT at predicting entities composed of multiple tokens (large entities). For example, for French HIPE, from 170 entities with a length equal or higher than five tokens[15], the stand-alone BERT managed to correctly detect 70% of them, while both BERT+n×Transf models

[15] The length of French HIPE entities ranges from one to 21 tokens.

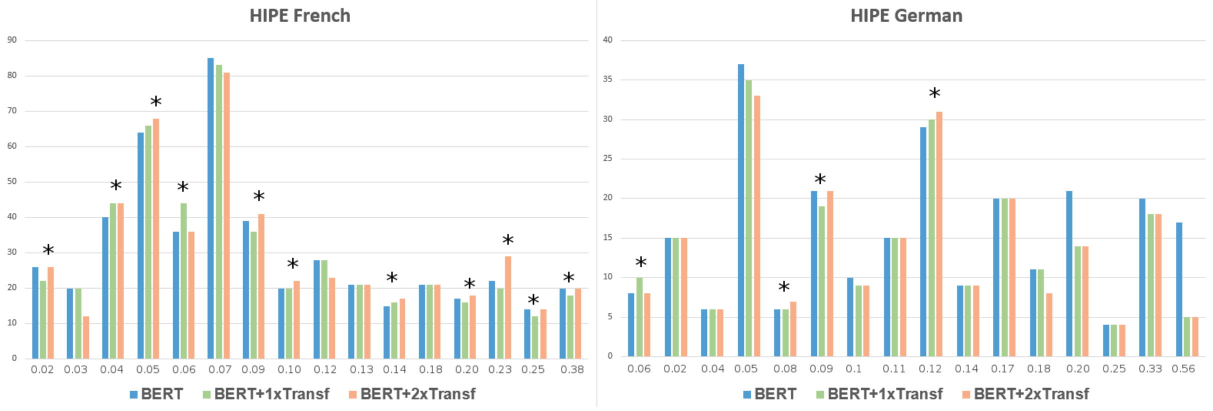

Figure 3: Correct predictions of misspelled entities based on the Levenshtein Ratio.

correctly identified 72.94% of them. German HIPE has less entities longer than five tokens[16], more exactly 97, and while the stand-alone BERT detected 50.51% of them, the BERT+$n\times$Transf models correctly detected and classified 55.67% for $n = 1$ and 54.63% for $n = 2$. In the following examples from Table 4, our method correctly predicted the full entity frequently while the stand-alone BERT only predicted a part of it.

Analyzing the French predictions for BERT and BERT+$n\times$Transf, we observed that BERT detects on average 75.04% of the entities of size 1 to 10, with other models performing slightly better. However, for entities with more than 10 tokens, there is clear a difference, since BERT detects 55.54% of the entities, while BERT+1×Transf detects 57.13%, and BERT+2×Transf reaches 82.52%. Examples are given in Table 4.

Gold standard	Predicted by	
	BERT	**BERT+$n\times$Transf**
signéKocH, avocat	, avocat	signéKocH**, avocat**
district de Gumbinnen	Gumbinnen	district de **Gumbinnen**
Armel Guerne. son adaptateur	Armel Guerne	**Armel Guerne**. son adaptateur
M. Javits, sénateur de New York juif et pro- israèlien	M. Javits, sénateur de New York	**M. Javits, sénateur de New York** juif et pro- israèlien

Table 4: Examples of long entities predicted by all models (the entity parts detected by BERT alone are highlighted in bold font under BERT+$n\times$Transf).

In the lower part of Figure 2, we present a German example where BERT becomes confused and predicts multiple partial spurious entities in a sentence. One can also observe that these entities are of two of the most common types in the dataset, persons (PERS) and locations (LOC). In this case, there is an overprediction of these types, which leads us to the interpretation that BERT is sensitive to misspellings and might overfit on OCR-related patterns. This observation proves that BERT has unbalanced attention to misspelled or corrupted words when the most informative words contain such errors (Sun et al., 2020).

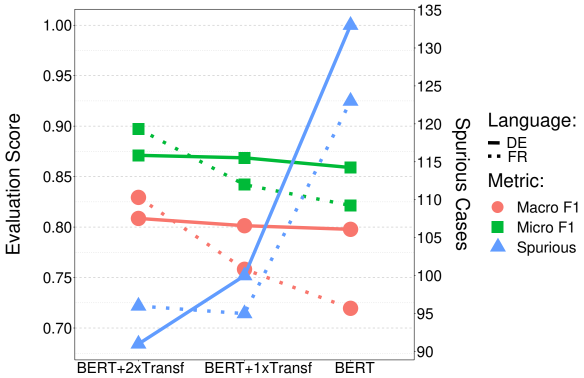

Figure 4: Number of spurious entities with respect to *micro-fuzzy* and *macro-fuzzy* F1 regarding the HIPE corpus.

To assess these assumptions, in Figure 4, we compare, per model and language, the values of *micro-fuzzy* F1 and *macro-fuzzy* F1 in the HIPE corpus. We include, as well, the number of spurious cases, i.e. tokens that were considered as an entity, despite not belonging to one, such as '*Zusammenziehung*' in Figure 2.[17] Due to the difference between *micro* and *macro* metrics, we can

[16] The length of German HIPE entities ranges from one to 16 tokens.

[17] We obtained the spurious cases by searching for predicted named entities that did not correspond, partially or totally, to one in the gold standard.

ascertain that the three presented models focused on predicting the most frequent entity types, i.e. PERS and LOC. Moreover, we can see that BERT achieved its result by creating more spurious cases in comparison to BERT+$n\times$Transf. This could mean that BERT learned that overpredicting was a straightforward solution to achieve better results. In the case of BERT+$n\times$Transf, we can see that the Transformer layers made the models to be more conservative and at the same time more accurate in their predictions.

7 Conclusions and Future Work

We presented a deep learning architecture for NER based on stacked Transformer layers that includes a fine-tuned BERT encoder and several Transformer blocks. Results on two historical datasets in French and German showed the fitness of the proposed model to process noisy digitized text corpora in distinct languages. At the same time, the approach did not degrade the performance over modern data. Thus, this type of model appears to be adapted for the NER of historical document collections.

While the improvements brought by the proposed NER model are clear, our analysis of the results highlighted several factors that could influence the results. Further analysis remains to be done. Thus, hereafter, we will investigate detailed variations of our architecture. In addition, we intend to explore data augmentation techniques, simulating digitized data by adding noise to digitally-born documents. This could be a solution to increase the size and expand the diversity of training datasets for performing NLP tasks over historical documents.

Acknowledgments

This work has been supported by the European Union's Horizon 2020 research and innovation program under grants 770299 (NewsEye) and 825153 (Embeddia).

References

Gustavo Aguilar, Suraj Maharjan, Adrian Pastor López-Monroy, and Thamar Solorio. 2017. A multi-task approach for named entity recognition in social media data. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 148–153, Copenhagen, Denmark. Association for Computational Linguistics.

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.

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.

Alexei Baevski, Sergey Edunov, Yinhan Liu, Luke Zettlemoyer, and Michael Auli. 2019. Cloze-driven pretraining of self-attention networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5360–5369, Hong Kong, China. Association for Computational Linguistics.

Jason PC Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional lstm-cnns. *Transactions of the Association for Computational Linguistics*, 4:357–370.

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

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(ARTICLE):2493–2537.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 7059–7069. Curran Associates, Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, Minneapolis, Minnesota. Association for Computational Linguistics.

Maud Ehrmann, Matteo Romanello, Stefan Bircher, and Simon Clematide. 2020a. Introducing the clef 2020 hipe shared task: Named entity recognition and linking on historical newspapers. In *European Conference on Information Retrieval*, pages 524–532. Springer.

Maud Ehrmann, Matteo Romanello, Simon Clematide, Phillip Benjamin Ströbel, and Raphaël Barman. 2020b. Language resources for historical newspapers: the impresso collection. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 958–968.

Alex Erdmann, Christopher Brown, Brian D Joseph, Mark Janse, Petra Ajaka, Micha Elsner, and Marie-Catherine de Marneffe. 2016. Challenges and solutions for Latin named entity recognition. In *COLING 2016: 26th International Conference on Computational Linguistics*, pages 85–93. Association for Computational Linguistics.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1243–1252. JMLR.org.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Ahmed Hamdi, Axel Jean-Caurant, Nicolas Sidere, Mickaël Coustaty, and Antoine Doucet. 2019. An analysis of the performance of named entity recognition over ocred documents. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 333–334. IEEE.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Helena Hubková. 2019. *Named-entity recognition in Czech historical texts: Using a CNN-BiLSTM neural network model*. Ph.D. thesis.

Ioannis Kanaris, Konstantinos Kanaris, Ioannis Houvardas, and Efstathios Stamatatos. 2007. Words versus character n-grams for anti-spam filtering. *International Journal on Artificial Intelligence Tools*, 16(06):1047–1067.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Daniel Lopresti. 2009. Optical character recognition errors and their effects on natural language processing. *International Journal on Document Analysis and Recognition (IJDAR)*, 12(3):141–151.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

John Makhoul, Francis Kubala, Richard Schwartz, Ralph Weischedel, et al. 1999. Performance measures for information extraction. In *Proceedings of DARPA broadcast news workshop*, pages 249–252. Herndon, VA.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

David Miller, Sean Boisen, Richard Schwartz, Rebecca Stone, and Ralph Weischedel. 2000. Named entity extraction from noisy input: speech and ocr. In *Proceedings of the sixth conference on Applied natural language processing*, pages 316–324. Association for Computational Linguistics.

Benjamin Muller, Benoît Sagot, and Djamé Seddah. 2019. Enhancing bert for lexical normalization. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 297–306.

Lluís Padró and Evgeny Stanilovsky. 2012. FreeLing 3.0: Towards Wider Multilinguality. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 2473–2479, Istanbul, Turkey. ELRA.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2018)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Matthew E Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. *arXiv preprint arXiv:1705.00108*.

Danish Pruthi, Bhuwan Dhingra, and Zachary C. Lipton. 2019. Combating adversarial misspellings with robust word recognition. In *57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pages 5582–5591, Florence, Italy.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Martin Riedl and Sebastian Padó. 2018. A named entity recognition shootout for german. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 120–125.

Kepa Joseba Rodriquez, Mike Bryant, Tobias Blanke, and Magdalena Luszczynska. 2012. Comparison of named entity recognition tools for raw OCR text. In *11th Conference on Natural Language Processing, KONVENS 2012, Empirical Methods in Natural Language Processing, Vienna, Austria, September 19-21, 2012*, volume 5 of *Scientific series of the ÖGAI*, pages 410–414. ÖGAI, Wien, Österreich.

Cícero dos Santos and Victor Guimarães. 2015. Boosting named entity recognition with neural character embeddings. In *Proceedings of the Fifth Named Entity Workshop*, pages 25–33, Beijing, China. Association for Computational Linguistics.

Cícero dos Santos and Bianca Zadrozny. 2014. Learning character-level representations for part-of-speech tagging. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1818–1826.

Silke Scheible, Richard J. Whitt, Martin Durrell, and Paul Bennett. 2011. Evaluating an 'off-the-shelf' POS-tagger on early modern German text. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 19–23, Portland, OR, USA. Association for Computational Linguistics.

Stefan Schweter and Johannes Baiter. 2019. Towards robust named entity recognition for historic German. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 96–103, Florence, Italy. Association for Computational Linguistics.

Daniel van Strien, Kaspar Beelen, Mariona Coll Ardanuy, Kasra Hosseini, Barbara McGillivray, and Giovanni Colavizza. 2020. Assessing the impact of ocr quality on downstream nlp tasks.

Lichao Sun, Kazuma Hashimoto, Wenpeng Yin, Akari Asai, Jia Li, Philip Yu, and Caiming Xiong. 2020. Adv-bert: Bert is not robust on misspellings! generating nature adversarial samples on bert. *arXiv preprint arXiv:2003.04985*.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Vikas Yadav and Steven Bethard. 2018. A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2145–2158, Santa Fe, New Mexico, USA. Association for Computational Linguistics.