Transferring Modern Named Entity Recognition to the Historical Domain: How to Take the Step?

Baptiste Blouin, Benoit Favre, Jeremy Auguste, Christian Henriot

Transferring Modern Named Entity Recognition to the Historical Domain: How to Take the Step?

Baptiste Blouin[1,2], Benoit Favre[1], Jeremy Auguste[2], and Christian Henriot[2]

[1]Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France
firstname.lastname@lis-lab.fr
[2]Aix Marseille Univ, Institut de Recherches Asiatiques, Aix, France
firstname.lastname@univ-amu.fr

Abstract

Named entity recognition is of high interest to digital humanities, in particular when mining historical documents. Although the task is mature in the field of NLP, results of contemporary models are not satisfactory on challenging documents corresponding to out-of-domain genres, noisy OCR output, or old-variants of the target language. In this paper we study how model transfer methods, in the context of the aforementioned challenges, can improve historical named entity recognition according to how much effort is allocated to describing the target data, manually annotating small amounts of texts, or matching pre-training resources. In particular, we explore the situation where the class labels, as well as the quality of the documents to be processed, are different in the source and target domains. We perform extensive experiments with the transformer architecture on the LitBank and HIPE historical datasets, with different annotation schemes and character-level noise. They show that annotating 250 sentences can recover 93% of the full-data performance when models are pre-trained, that the choice of self-supervised and target-task pre-training data is crucial in the zero-shot setting, and that OCR errors can be handled by simulating noise on pre-training data and resorting to recent character-aware transformers.

1 Introduction

Due to the massive effort to digitize and transcribe historical documents, the field of digital humanities is facing the challenges of digesting and analyzing large quantities of texts. With the continuous advancements of natural language processing (NLP), there is a growing interest in applying tools such as named entity recognition (NER) on historical documents. Indeed, identifying people, places, and other historical entities is crucial to understand the historical context, and having the ability to do so automatically is a major step forward.

	Models	LitBank	HIPE
Off the shelf	Spacy, en_core_web_sm	21.27	12.35
	Spacy, en_core_web_trf	28.36	19.60
	Stanford CoreNLP	23.59	31.23
SOTA	Boros et al.	–	63.20
	Ju et al.	68.30*	–
Ours	Zero-shot	70.44	13.73
	Full	81.53	62.32

Table 1: Off-the-shelf NER performance on historical texts from the LitBank and HIPE test sets in a zero-shot setting and performance of State-Of-The-Art systems trained on target data. Reported values are F1-scores. * denotes a somewhat different experimental setting, which makes this result incomparable.

It facilitates the exploration of massive corpora by identifying, counting and extracting textual clues, among others to enrich a database, which can help to systematically explore the information reported in these documents. But the variety present in historical texts, compared to modern ones, makes the evaluation and application of NLP techniques quite difficult. In particular, apart from the fact that these documents relate to different domains, the evolution of language, as well as the noise due to optical character recognition (OCR) errors, preclude using off-the-shelf systems.

NER success, like for other NLP tasks, is highly dependent on the corpus on which the system has been trained, and most of the available named entity (NE) corpora use contemporary texts with contemporary concerns. For example, off-the-shelf NER systems such as SpaCy (Honnibal et al., 2020) or Stanford's CRF-NER (Manning et al., 2014) do not yield acceptable results on historical texts as evidenced in Table 1. Therefore, domain adaptation and the zero-shot setting are very relevant to applying NLP on historical documents. In this study, zero-shot setting denotes the training of a

system on contemporary data and an evaluation on historical data.

The goal of this study is to evaluate the effort required to obtain relevant NER results on historical documents. This effort can relate to annotation in the target domain, transfer of contemporary models, pre-training on matching resources, or adaptation to OCR errors. Compared to other domain transfer approaches (Jia et al., 2019; Liu et al., 2020; Sachan et al., 2018), where the evaluation is carried out on specific contemporary domains, this work considers NER from a historian's point of view: we wish to process historical texts and understand why some approaches do not yield usable results. In that context, would off-the-shelf systems be appropriate? If not, can we reduce the amount of annotation needed to obtain reasonable results by using existing resources? And finally, if so, is this approach robust to the characteristic difficulties of historical documents?

Our contribution is to tentatively answer the following questions: (1) What annotation effort in the target historical domain? (2) Is it worth adapting initial pre-trained word representations? (3) What is the impact of OCR errors on transfer performance? We perform experiments on the LitBank (Bamman et al., 2019) and HIPE (Ehrmann et al., 2020) annotated historical NER datasets using prototypical NER systems built on the previous and current generation of models, trained on contemporary annotations from ACE 2005 (Walker et al., 2006) and various amounts of target data.

2 Related work

NER is a typical sequence labeling task where the aim is to locate and classify named entities mentioned in unstructured text into pre-defined categories such as person names, organizations, locations, etc. The research around this task has allowed obtaining very good results on modern documents. For example, on the CoNLL-03 dataset the results reach up to 94.9% of F1-score. From the numerous existing approaches, we describe a few representative recent works. Wang et al. (2020b) propose an Automated Concatenation of Embeddings to build adequate system input representations for structured predictions. It will calculate error based on results of the training process and then compare it with other combinations to finally find out the most suitable concatenation embedding layer for the problem. Yamada et al. (2020) propose a model that treats words and entities in a given text as independent tokens, and outputs contextualized representations of them. Their Transformer based language model pre-trained on both large-scale text corpora and knowledge graphs achieves SOTA performance on various entity related tasks. Wang et al. (2021) propose to use external contexts to improve model performance by retrieving and selecting a set of semantically relevant texts through a search engine and constructed a new representation through the concatenation of a sentence and its external contexts.

Given the progress on this task and its need for applications, many off-the-shelf models pre-trained on modern data are made available. Spacy (Honnibal et al., 2020), Flair (Akbik et al., 2019), Stanza (Qi et al., 2020), AllenNLP (Gardner et al., 2018) offer models trained on OntoNotes 5.0, Stanford CoreNLP (Manning et al., 2014) provides a model trained on a mixture of CoNLL, MUC-6, MUC-7 and ACE.

In this paper we mainly focus on NER in English but this task is also a subject of research in other languages. For instance, Cao et al. (2018) proposed an adversarial transfer learning strategy to make full use of the boundary information shared by tasks and prevent the task-specific functions of Chinese word segmentation. (Rahimi et al., 2019) process 41 languages using truth inference to model the transfer annotation bias from diverse source-languages models. Xie et al. (2018) created a 0-shot NER systems by aligning monolingual embeddings from English to Spanish, German, and Dutch, and then translated the English CoNLL dataset into these languages, and built a self-attentive Bi-LSTM-CRF model using the translated languages.

Cross domain NER Cross-domain approaches have been developed to enhance the generalization of NER models to a given target domain.

Most existing approaches are in a supervised setting where both source and target domains have labeled data, the goal being to improve performance over only using instances from the target data. Baselines jointly train models on source and target data with shared parameters, add adaptation layers on top of source models for capturing target domain specifics, or design label-aware feature representations for NER adaptation (Daumé III, 2007; Wang et al., 2018a).

More specific methods use multi-task ap-

proaches which have been shown to be effective for this cross-domain task, to reduce the gap between the two domains. For example, Jia et al. (2019) propose to use extrinsic data in both the source and target domains to train language models for domain adaptation. Wang et al. (2018b) propose a parameter transfer learning between feature representations from Bi-LSTM and two conditional random fields. Wang et al. (2020a) propose a multi-task learning objective that learns domain labels as an auxiliary task and Zhou et al. (2019) propose a Dual Adversarial Transfer Network which aims at addressing representation difference and resource data imbalance problems.

Methods such as transfer learning also show that knowledge sharing is effective for cross-NER. For instance, Lee et al. (2018) utilize transfer learning by initializing a target model with parameters learned from source-domain NER, and rely on labeled target domain data to finetune the model. Cao et al. (2018) propose an adversarial transfer learning framework for Chinese NER task, which can exploit task-shared word boundaries features and ensure proper information usage from the word segmentation task. Lin and Lu (2018) perform adaptation across two domains using adaptation layers augmented on top of the existing neural model. Yang et al. (2019) propose a fine-grained knowledge fusion model to balance the learning from the target data and learning from the source model.

NER on historical texts Given the number of general NER papers produced over the last few decades in the NLP field, studies targeting historical texts and literary documents are still scarce.

In general, research on this subject have not only experimented with NER applied to historical materials but also many of them have addressed some of the most pressing challenges involved in the use of current state-of-the-art NER systems on historical texts: disparate quality of digitization and OCR, handling of non-European or classical languages, or dealing with spelling variations. Packer et al. (2010) experimented with recognition of person names in noisy OCR texts using a Dictionary-Based, Regex-Based, Maximum Entropy Markov and CRF models, and evaluated the output against hand-labelled test data. Grover et al. (2008) built a rule-based NER system for recognizing names of places and persons in digitized records by focusing on issues caused by the high level of variance in the use of word-initial upper-case letters, as well as issues connected to the use of OCR technology. Rodriguez et al. (2012) evaluated four tools for NER on historical texts including OpenNLP (Kwartler, 2017) and Stanford NER. They showed that the Stanford NER system had the overall best performance. Rodrigues Alves et al. (2018) show that character-level word embedding, combined with a Bi-LSTM-CRF model, can help reduce the impact of OCR errors and handle rare words in 19-21C scholarly books and journals. More recent approaches (Schweter and März, 2020) evaluated the impact of word embeddings at the level of their learning and their combination, on this task. Labusch et al. (2019) apply a model based on multilingual BERT embeddings, which is further pre-trained on large OCRed historical German unlabelled data and subsequently finetuned on several NER datasets. They show that an appropriately pre-trained BERT model delivers decent performance in a variety of settings. Boros et al. (2020) added a two task-specific transformer layers on top of the pre-trained BERT to alleviate data sparsity issues. However, the use of recent word representations, such as BERT, is not totally suitable, as its ability to handle noisy data remains a point to be clarified as to its robustness (Sun et al., 2020).

Compared to them, we do not seek to optimize the performance on specific historical data, but rather propose a replicable transfer procedure linking the effort to be provided on the target domain in order to have performance relative to those obtained on contemporary data.

3 Datasets

In this study, we deal with target domain annotated datasets (historical texts), and source domain annotated datasets (contemporary texts, typically news). Table 2 outlines dataset statistics for both domains.

Two target datasets are used, each with two different subdomains, specific difficulties towards this task and a non-comparable annotation guideline.

LitBank (Bamman et al., 2019) is an annotated dataset of 100 English-language literary public-domain texts from Project Gutenberg, annotated with ACE entity categories except for the weapon category (person, location, geopolitical entity (GPE), facility, organization (ORG), and vehicle). In contrast to existing datasets built primarily on news (focused on GPEs and ORGs), literary texts offer strikingly different distributions of entity categories, with much stronger emphasis on

Domain	Dataset	Train	Dev	Test	NE freq.	Types	Text sources	Time Period
Target	LitBank	29,894	4,133	3,425	18%	7	Novels	1852-1923
	HIPE	0	2,575	1,301	9%	5	Newspapers	1798-2018
Source	ACE 2005	34,669	4,336	3,777	24%	7	News, speech, web	2003-2004

Table 2: Datasets statistics. **train/dev/test** columns represent the number of named entities. **NE freq.** represents the ratio between the number of entities and the number of words. **Types** indicates the number of different entity categories.

people and description of settings. All texts were published before 1923, with the majority falling between 1852 and 1911.

HIPE (Ehrmann et al., 2020) is a collection of digitized documents covering three different languages: English, French, and German. The documents come from archives of several Swiss, Luxembourgish, and American newspapers. The corpus was manually annotated by native speakers according to the HIPE impresso guidelines, which are derived from the Quaero[1] annotation guide. The corpus is annotated with 5 types of entities: person, location, organization, time and production. The time-span of the whole corpus goes from 1798 to 2018. The particularity of this dataset is that it contains OCR errors, with no gold alignment. Feuilleton, tabular data, crosswords, weather forecasts, time schedules and obituaries were excluded as well as articles that were fully illegible due to OCR errors. In this study, only the English part of this corpus is used: annotations are only available for development and testing, but not for training.

One source dataset is used. **ACE 2005** (Walker et al., 2006) Multilingual Training Corpus was developed by the Linguistic Data Consortium (LDC) and contains approximately 1,800 documents of mixed genre texts in English, Arabic, and Chinese annotated for entities, relations, and events. The genres include newswire, broadcast news, broadcast conversation, blog, discussion forums, and conversational telephone speech. The dataset is annotated with 7 entity types: person, location, GPE, facility, organization, vehicle, and weapon. We followed the same pre-processing of the data as those presented by Bamman et al. (2019).

In general, all corpora share annotations in persons, locations and organizations, these three types also represent the majority of annotations. ACE and LitBank share 100% of their entity types and ACE and HIPE share 3 out of 6 entity types, which represents 92.5% of entity instances of shared type in the test set. A shared entity type is an entity type with the same name (e.g., Person).

4 Experimental Settings

We used three different systems which represent the typical kind of systems that could be implemented in an industry-provided solution for historical NER given current technology.

BERT (Devlin et al., 2019), with an extra layer to predict NER categories. The pre-trained contextual embeddings are finetuned on the source/target dataset using their proposed approach. It's a *de facto* baseline for NLP systems, and is implemented with the transformers library (Wolf et al., 2020). We also used BERT models (**BERT-Hist**) (Hosseini et al., 2021) finetuned on 47,685 books (5.1B tokens) in English from the year 1760 to 1900 from the Microsoft British Library Corpus. This model, compared to BERT-Base learned from Wikipedia and Bookcorpus, allows us to question the impact of diachronic language changes on these embeddings applied to the historical domain.

CharBERT (Ma et al., 2020) which accounts for characters in addition to BPE tokens. The model is pre-trained with a Noisy LM objective for obtaining robust character-level representations. We expect such model to perform well on noisy OCR, and rely on the available implementation[2].

LSTM-CRF (Lample et al., 2016) initialized with FastText (Bojanowski et al., 2016) non-contextual embeddings, an architecture that has shown its robustness for NER and is standard in many available systems. This allows to compare its performance to contextual ones. The Flair library (Akbik et al., 2019) is used for this model.

All results presented in the experiments section are averages over 10 random initializations. Since the objective is not to maximize performance on a particular dataset or on a particular architecture, the same hyperparameters for all experiments of a given system are used.

The learning rate is initialized to 0.1 for the

[1] http://www.quaero.org/media/files/bibliographie/quaero-guide-annotation-2011.pdf

[2] https://github.com/wtma/CharBERT

LSTM-CRF and to 3e-5 for the transformers. For the three architectures, the size of the hidden layers are set to 512 and the training is performed on 3 epochs on the source data (10 epochs for the LSTM-CRF) and finetuned on 10 epochs on the target data. The rest of the parameters are the defaults proposed by the libraries for the different models. NER performance is computed with F1-score (F1). Because not all corpora contain nested entities, the embedded entity mention is removed and the smaller mention is kept in the case where the datasets proposed this type of entity. Moreover, as previously mentioned, the HIPE dataset do not share all entity types. The evaluation of a system is only performed on the entity types of the test set, and all the predicted entities that are not part of the test set tagset are removed prior to evaluation.

5 Experiments

5.1 What annotation effort in the target domain?

The low performance of off-the-shelf systems, exemplified in Table 1, suggests that adaptation is necessary.

First, systems are evaluated according to the amount of annotated data in the target domain. For this experiment, the ACE English NER dataset is considered as the source domain, and LitBank and HIPE as the target domain. LitBank originates from the same annotation guidelines as ACE while HIPE is based on different guidelines. We first pretrain the different systems on the source domain data, then we finetune them on the target domain samples with a varying quantity of inputs, from 10 sentences to the maximum number of sentences available in the target corpus, up to 6k sentences for LitBank, which is already a large number of annotated sentences. We compare this approach to a system trained only on the target data. This experiment allows evaluating the expected performance given the annotation effort in the target domain, and outlines the importance of pre-training. The results of this experiment are given in Table 3.

First, when the systems are already pre-trained, depending on the amount of target data used, the results, independently of the system used, vary from 17.7% to 27.3% of F1 using 10 sentences and from 46.8% to 61.1% of F1 using the whole HIPE corpus. In the case where our systems were only trained on the HIPE dataset, the three systems obtain 0% F1 for a training on 10 sentences and can vary from 39.1% to 57.6% using the whole corpus.

When pre-trained on ACE data, the three systems present similar score evolutions according to the amount of data used (modulo the maximum value obtained by each system). Only 10 sentences of the target dataset are required to achieve performance that is more than one third of the maximum performance that could be achieved using the entire dataset. Compared to training only on 10 sentences, without pre-training, where the systems fail to learn. By using 50 sentences from the target dataset one can obtain more than two thirds of the maximum performance, so compared to training without pre-training on source data, 50 sentences is still insufficient to generalize on the test set, except for CharBERT, which in this case manages to get more than half of the maximum performance. Above this quantity of annotated sentences, the finetuning approach presents a constant improvement while keeping results superior to training from scratch on the target data. However, from 250 annotated sentences BERT and LSTM-CRF have enough data to learn without ACE data, at this stage, these two systems can get about 75% of the maximum performance. Concerning CharBERT, it recovers 95% of the maximum performance using about half of the available data.

The results obtained on LitBank are similar to those observed on HIPE except that in the case of finetuning from ACE, the target dataset shares the same annotation guideline as the source and doesn't contain noise, which explains the high performance even with low amounts of target data.

However, since this dataset provides more annotated sentences, the analysis can be taken further. At 400 sentences on LitBank, in the case of a pre-training on ACE, systems are within 92-96% of the maximum performance when using 6000 sentences. But in the case of training from scratch on the target, performance is well below with 86%, 79% and 61% of the maximum F1 for CharBERT, BERT and LSTM-CRF respectively. When using 6000 sentences, except for LSTM-CRF, not using pre-training gives the same results as when using it. In practice, having 6000 annotated sentences already requires a big annotation effort. In these results, transfer approaches show that they require fewer annotated sentences than training from scratch from a more realistic amount of 1000 annotated sentences. Indeed, for BERT and LSTM-CRF, in the case of pre-training on ACE and with the addition of 250

Models / Splits		LitBank							HIPE				
		10	50	250	400	1000	3000	6000	10	50	250	400	444
Pre-trained	CharBERT	68.1	71.0	75.1	76.0	77.6	79.8	80.6	27.3	46.9	57.9	61.1	61.1
	BERT	69.4	73.3	75.9	76.8	78.9	80.7	80.9	25.2	47.8	54.6	57.4	58.1
	LSTM-CRF	55.9	62.4	67.6	69.4	71.5	74.7	75.5	17.7	31.6	44.1	46.8	46.8
Only	CharBERT	00.0	30.5	64.0	69.2	74.5	78.8	80.1	00.0	33.8	54.2	57.3	57.6
	BERT	00.0	00.0	54.3	63.7	72.9	79.0	80.4	00.0	00.0	42.6	51.5	52.1
	LSTM-CRF	00.0	00.0	24.5	44.4	57.8	68.3	72.9	00.0	00.4	29.6	37.6	39.1

Table 3: F1 obtained on the targets test set depending on the system used as well as the amount of training on the target used in the case where our systems are already **pre-trained** on ACE and in the case where our systems were **only** trained on the target.

annotated sentences on the target corpus, better performances are obtained (75.9% and 67.6% of F1 respectively) than using models trained from scratch on 1000 sentences (72.9% and 57.8% of F1 respectively). CharBERT requires 400 annotated sentences when it is pre-trained on ACE to obtain better performance than a system learned only on 1000 sentences from LitBank. This increase can be explained by the fact that CharBERT requires less data than other systems to learn on new data. Indeed, 50 annotated sentences allows obtaining 30.5% of F1 with CharBERT compared to 0% with the other systems.

Due to the different annotation guidelines, the type of data and the quality of the documents used, cross-NER to a historical domain requires at least some annotation in the target domain. Nevertheless, we could see that pre-training a system on contemporary data allows to considerably decrease the amount of annotation needed. Through these experiments two thresholds are observed, the first one at 250 sentences, which allows obtaining very promising results on a distant domain on a low budget. The second, estimated at 1000 sentences, allows obtaining almost optimal performance.

5.2 Is it worth adapting initial word representations?

A realistic scenario for digital humanities is to have access to large annotated corpora in the target domain. However, reaching the scale of data required for pre-training a BERT-like language model is unlikely in the target domain and training such models from scratch is still computationally expensive. Jia et al. (2019) show that finetuning contextual embeddings with pre-training task on a relatively small amount of unannotated texts can improve transfer results. The approach, called domain adaptive pre-training (DAPT), consists in adapting word representations prior to training and transferring the NLP task at hand. Therefore, we evaluate the impact of finetuning the BERT representations to the target texts prior to training the NER system compared to using a BERT trained on historical data (Hosseini et al., 2021). The experimental procedure is described in Figure 1.

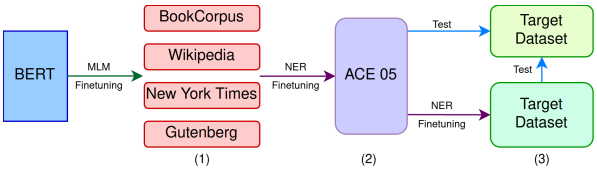

Figure 1: A three-step procedure: (1) BERT representations are first finetuned on two million sentences from unannotated corpora (BookCorpus fiction, Wikipedia, New York Times newspaper, or Gutenberg fiction), (2) the representations are further finetuned to train an ACE NER system thanks to an added decision layer, and (3) that model is finally finetuned on the target NER annotated dataset.

The four pre-training datasets have been selected for the following reasons. **Perfect match**: Gutenberg is the source corpus for LitBank, it represents the perfect adaptation corpus due to its proximity to the target. **Genre match, time mismatch**: The New York Times corpus represents a partial match since, like HIPE, it is a newspaper corpus, but it is not from the same period nor from the same source. **Similar genre and time**: By finetuning the representations on Bookcorpus we want to focus the representations on its literary side in order to observe the improvement obtained on LitBank, without totally modifying the distribution following the addition of new data. **Complete mismatch**: Wikipedia does not share anything with the target but is a general domain corpus.

As a comparison, we add randomly initialized BERT baselines to question the importance of the match between the pre-training domain and the source and target domains. Finally, we finetuned a BERT only on the NER training data of the target.

The results of this experiment (table 4 for Lit-

Pre-training	LitBank		
	Only	*0-shot*	*Full*
No init baseline	34.00	10.71	34.02
BERT-Base	80.40	**70.44**	80.80
BERT-Hist	79.76	63.91	80.09
Book Corpus	79.60	66.66	79.63
Wikipedia	79.83	67.01	80.20
New York Times	78.69	65.13	79.75
Gutenberg	**81.03**	64.64	**81.53**
LitBank-NER	80.53	70.31	80.41

Table 4: NER F1 performance obtained on the LitBank test sets according the unannotated corpus on which BERT pre-training tasks are finetuned. A zero-shot scenario (only trained ACE), a full training scenario (trained ACE then finetuned on target data) and in the case where our systems were **only** trained on the target are compared.

Pre-training	HIPE		
	Only	*0-shot*	*Full*
No init baseline	08.34	01.72	08.44
BERT-Base	52.10	13.44	58.14
BERT-Hist	**60.96**	09.61	**58.34**
Book Corpus	51.63	11.60	53.51
Wikipedia	52.98	**13.73**	56.19
New York Times	51.26	11.96	54.88
Gutenberg	56.41	12.97	57.86
HIPE-NER	53.50	09.84	55.14

Table 5: NER F1 performance obtained on the HIPE test sets according the unannotated corpus on which BERT pre-training tasks are finetuned. A zero-shot scenario (only trained ACE), a full training scenario (trained ACE then finetuned on target data) and in the case where our systems were **only** trained on the target are compared.

Bank and table 5 for HIPE) show that finetuning word representations has a small, often negative, impact on NER performance. In the full training scenario, using DAPT improves NER in the LitBank case (+0.78 compared to BERT) by using matching data (Gutenberg). Surprisingly, that same data corresponds to the worst results for zero-shot on LitBank (-5.79 points). On HIPE, where no target training data are available, BERT-base is the most stable in the transfer scenarios and performance is very low in the 0-shot setting across the board, due to mismatch in annotation guidelines. In addition, for this data set, being distant in annotation and domain from our source data, learning using only the target data with historical embeddings leads to better results (60.96).

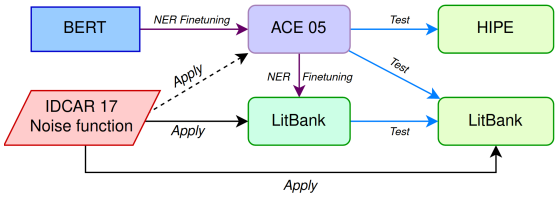

Figure 2: Procedure used to evaluate the impact of noise on the transfer models. The dotted arrow corresponds to the second part of the experiments.

5.3 What is the impact of OCR errors on transfer performance?

Historical texts are often the result of digitization and OCR from the original paper documents. This process leads to inevitable errors due to the quality of the source material and the variability of the print or handwriting.

The last experiment presented in this paper evaluates the impact of OCR errors on NER. Given the target corpora, it's difficult to evaluate their real impact. On the one hand the HIPE texts contain OCR errors, but we do not have manually corrected texts, and on the other hand, the LitBank texts are OCR-error free. Therefore, we chose to simulate an OCR system by randomly adding errors to LitBank.

Following previous work (Pruthi et al., 2019), we change the original character sequence by deleting, inserting, and substituting characters within it. Figure 2 illustrates the procedure that is used in order to mimic OCR errors and to evaluate their influence on the NER task. In order to obtain a realistic error distribution, the error probabilities are computed from the ICDAR 17 corpus, which contains both OCR from several systems and reference texts. This approach allows to control the amount of noise applied to the dataset, simulating various OCR difficulty settings. As a first step, transfer quality from regular ACE to noisy LitBank is evaluated with BERT, BERT-Hist and CharBERT systems, the assumption being that the latter shall cope better with errors. The results as a function of the amount of noise injected into the target corpus are shown in table 6. In a second step, the same transfer is evaluated after corrupting the source data (ACE) using the same process. We varied the amount of noise from 0% to 10% calculated as the character error rate. Results are presented in tables 7 and 8.

Obviously, the two systems are sensitive to noise, and as the amount of noise increases, F1 decreases. First, it can be seen in Table 6 that combining a

Noise	Only			0-shot			Full		
	BERT-Hist	*BERT*	*CharBERT*	*BERT-Hist*	*BERT*	*CharBERT*	*BERT-Hist*	*BERT*	*CharBERT*
0%	79.74	80.40	80.10	63.91	70.44	67.38	80.08	80.80	80.61
1%	79.01	78.46	79.46	64.37	67.73	66.17	79.09	78.96	79.92
2.5%	76.28	75.69	77.72	61.61	64.03	63.86	76.78	76.45	78.14
5%	73.27	71.46	75.36	58.29	57.81	60.58	73.93	71.97	75.79
10%	67.00	63.83	70.52	47.71	47.76	48.50	67.87	65.20	70.39

Table 6: Results obtained on the LitBank test depending on the amount of noise injected in the **target** corpus in a **0-shot** scenario (trained on ACE), in a **full** training scenario (trained on ACE, finetuned on LitBank) and when our systems were **only** trained on LitBank.

Noise	0-shot		
	BERT-Hist	*BERT*	*CharBERT*
0%	65.67	70.44	67.38
1%	64.95	68.03	66.76
2.5%	62.46	65.23	65.52
5%	59.15	60.73	62.25
10%	52.80	51.70	57.63

Table 7: Results obtained on the LitBank test depending on the amount of noise injected in the **source and target** corpora in a **0-shot** scenario (trained on ACE).

Noise	Full		
	BERT-Hist	*BERT*	*CharBERT*
0%	80.08	80.80	80.61
1%	79.37	78.69	79.54
2.5%	76.88	76.06	77.60
5%	73.72	72.14	75.50
10%	68.33	65.34	70.69

Table 8: Results obtained on the LitBank test depending on the amount of noise injected in the **source and target** corpora in a **full** training scenario.

system learned on clean data with noisy test data, CharBERT is less sensitive to noise compared to BERT the more degraded the data are. And the same goes for BERT-Hist which was learned from text created by OCR. In the case where we evaluate only the systems learned on ACE in a zero-shot scenario, the two systems show different behavior regarding noise. When the target data are noisy at 10%, we observe a strong drop of the F1, due to a drop in recall for BERT and BERT-Hist and to a drop in precision for CharBERT. In general, the results show that CharBERT is more robust to noise but suffers from the same performance degradation as BERT when only trained on clean data, but evaluated on noisy data. However, when a model first trained on clean data is finetuned with noisy texts, CharBERT's performance is much better than BERT's with a difference of 5.19 points of F1 at the 10% noise level. This difference, reduced to 2.52 points of F1 with BERT-Hist.

Compared to the previous results, when the same noise distribution is applied to the source dataset (Tables 7 and 8), performance is similar when finetuning on the noisy target dataset. However, in the case where the models trained on noisy ACE are directly evaluated on noisy LitBank, a large improvement is observed compared to when ACE is not noisy. Indeed, recall for the BERT models improves compared to when the source data was clean. The same behavior is observed for CharBERT, where also the recall improves.

However, in the zero-shot case, BERT learned as well as CharBERT up to 2.5% noise. Above that, the noisier the data are, the more robust CharBERT is compared to BERT, with a performance improvement of 5.93 points of F1 with 10% noise. Reduced to 4.83 points of F1 with BERT-Hist.

In view of the LitBank results with simulated noise, a similar approach can be applied for the source data when processing the HIPE corpus. We finetune CharBERT on ACE with 10% noise from the IDCAR 17 distribution and then finetune this model on HIPE which has its own natural OCR error distribution. The results show that training on noisy data, even though the noise follows a different error distribution, improves the results (+1.22 points) on a noisy dataset. However, in the zero-shot scenario, we have a drop of 0.56 points.

6 Discussion

We identify replicable steps for the application of NER on historical documents. First, select contemporary resources, pre-trained models or annotated datasets, using a guideline close to the target needs. Then annotate a few sentences on the target data to adapt the learned models on the source data. 250 sentences can recover 93% of the full-data performance and 1000 sentences 98%. In a second step, the choice of the model will depend on the quality of the target documents to process. If the target is noisy, CharBERT, even compared to a BERT learned on historical noisy data, is more robust to process this type of documents. Moreover, if the

quality information of the target documents is available, applying noise following the same distribution to the source documents will allow the system to be more robust on the target data. Finally, DAPT can be applied if a large amount of unannotated target data is available ($>$ 2M sentences) and if some target data is annotated. This approach does not seem to work in the case where no annotated target data is available. In the case of 0-shot, using word embeddings adapted to the source data will bring better performance on the target data.

7 Conclusion

In this work, we investigate the potential transfer of contemporary named entity recognition models to the historical domain. Experiments show that finetuning contemporary pre-trained transformers allows reducing considerably the annotation effort and can be further reduced by making an informed choice of the data sources for transfer. Adapting pre-trained word representations prior to learning the task (DAPT) allows a low-cost adaptation to the target domain and improves performance in the full settings depending on the dataset. Processing noisy data is still challenging but the choice of an architecture relying on pre-trained character representations, and the simulation of target noise on the source domain allows recovering acceptable performance compared to a BERT baseline.

To further this study, we would like to systematically look into recent transformer architectures and pre-train them on large corpora of historical texts instead of crawled web data. It would be an opportunity to infuse the models with knowledge of the target temporal span. In addition, we would like to study how levels of NER performance impact historian's findings, and whether current technology is acceptable for reliably mining large quantities of historical documents.

7.1 Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 788476).

This work was granted access to the HPC resources of IDRIS under the allocation 2021-AD011012594 made by GENCI.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. Flair: An easy-to-use framework for state-of-the-art nlp. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

David Bamman, Sejal Popat, and Sheng Shen. 2019. An annotated dataset of literary entities. In *Proceedings of the 2019 Conference of the North*, pages 2138–2144, Minneapolis, Minnesota. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomás Mikolov. 2016. Enriching word vectors with subword information. *CoRR*, abs/1607.04606.

Emanuela Boros, Elvys Linhares Pontes, Luis Adrián Cabrera-Diego, Ahmed Hamdi, José G. Moreno, Nicolas Sidère, and Antoine Doucet. 2020. Robust Named Entity Recognition and Linking on Historical Multilingual Documents. In *Conference and Labs of the Evaluation Forum (CLEF 2020)*, volume 2696 of *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum*, pages 1–17, Thessaloniki, Greece. CEUR-WS Working Notes.

Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, and Shengping Liu. 2018. Adversarial Transfer Learning for Chinese Named Entity Recognition with Self-Attention Mechanism. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 182–192, Brussels, Belgium. Association for Computational Linguistics.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic. Association for Computational Linguistics.

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, Alex Flückiger, and Simon Clematide. 2020. Extended Overview of CLEF HIPE 2020: Named Entity Processing on Historical Newspapers. In *CLEF 2020 Working Notes. Conference and Labs of the Evaluation Forum*, volume 2696. CEUR-WS. Number: CONF.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for*

NLP Open Source Software (NLP-OSS), pages 1–6, Melbourne, Australia. Association for Computational Linguistics.

Claire Grover, Sharon Givon, Richard Tobin, and Julian Ball. 2008. Named entity recognition for digitised historical texts. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Kasra Hosseini, Kaspar Beelen, Giovanni Colavizza, and Mariona Coll Ardanuy. 2021. Neural language models for nineteenth-century english. *CoRR*, abs/2105.11321.

Chen Jia, Xiaobo Liang, and Yue Zhang. 2019. Cross-Domain NER using Cross-Domain Language Modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2464–2474, Florence, Italy. Association for Computational Linguistics.

Meizhi Ju, Makoto Miwa, and Sophia Ananiadou. 2018. A neural layered model for nested named entity recognition. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1446–1459, New Orleans, Louisiana. Association for Computational Linguistics.

Ted Kwartler. 2017. *The OpenNLP Project*, chapter 8. John Wiley and Sons, Ltd.

Kai Labusch, Clemens Neudecker, and David Zellhöfer. 2019. Bert for named entity recognition in contemporary and historic german. In *KONVENS*.

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.

Ji Young Lee, Franck Dernoncourt, and Peter Szolovits. 2018. Transfer learning for named-entity recognition with neural networks. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Bill Yuchen Lin and Wei Lu. 2018. Neural Adaptation Layers for Cross-domain Named Entity Recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2012–2022, Brussels, Belgium. Association for Computational Linguistics.

Zihan Liu, Genta Indra Winata, Peng Xu, and Pascale Fung. 2020. Coach: A Coarse-to-Fine Approach for Cross-domain Slot Filling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 19–25, Online. Association for Computational Linguistics.

Wentao Ma, Yiming Cui, Chenglei Si, Ting Liu, Shijin Wang, and Guoping Hu. 2020. CharBERT: Character-aware pre-trained language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 39–50, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Thomas L. Packer, Joshua F. Lutes, Aaron P. Stewart, David W. Embley, Eric K. Ringger, Kevin D. Seppi, and Lee S. Jensen. 2010. Extracting person names from diverse and noisy ocr text. In *Proceedings of the Fourth Workshop on Analytics for Noisy Unstructured Text Data*, AND '10, page 19–26, New York, NY, USA. Association for Computing Machinery.

Danish Pruthi, Bhuwan Dhingra, and Zachary C. Lipton. 2019. Combating Adversarial Misspellings with Robust Word Recognition. *arXiv:1905.11268 [cs]*. ArXiv: 1905.11268.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Multilingual NER transfer for low-resource languages. *CoRR*, abs/1902.00193.

Danny Rodrigues Alves, Giovanni Colavizza, and Frédéric Kaplan. 2018. Deep reference mining from scholarly literature in the arts and humanities. *Frontiers in Research Metrics and Analytics*, 3:21.

Kepa J. Rodriguez, Mike Bryant, Tobias Blanke, and Magdalena Luszczynska. 2012. Comparison of named entity recognition tools for raw ocr text.

Devendra Singh Sachan, Pengtao Xie, Mrinmaya Sachan, and Eric P. Xing. 2018. Effective Use of Bidirectional Language Modeling for Transfer Learning in Biomedical Named Entity Recognition. *arXiv:1711.07908 [cs]*. ArXiv: 1711.07908.

Stefan Schweter and Luisa März. 2020. Triple E - effective ensembling of embeddings and language models for NER of historical german. 2696.

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:2003.04985 [cs]*. ArXiv: 2003.04985.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. ACE 2005 Multilingual Training Corpus. Type: dataset.

Jing Wang, Mayank Kulkarni, and Daniel Preotiuc-Pietro. 2020a. Multi-Domain Named Entity Recognition with Genre-Aware and Agnostic Inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8476–8488, Online. Association for Computational Linguistics.

Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Kewei Tu. 2020b. Automated concatenation of embeddings for structured prediction. *CoRR*, abs/2010.05006.

Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Kewei Tu. 2021. Improving named entity recognition by external context retrieving and cooperative learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1800–1812, Online. Association for Computational Linguistics.

Zhenghui Wang, Yanru Qu, Liheng Chen, Jian Shen, Weinan Zhang, Shaodian Zhang, Yimei Gao, Gen Gu, Ken Chen, and Yong Yu. 2018a. Label-aware double transfer learning for cross-specialty medical named entity recognition. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1–15, New Orleans, Louisiana. Association for Computational Linguistics.

Zhenghui Wang, Yanru Qu, Liheng Chen, Jian Shen, Weinan Zhang, Shaodian Zhang, Yimei Gao, Gen Gu, Ken Chen, and Yong Yu. 2018b. Label-Aware Double Transfer Learning for Cross-Specialty Medical Named Entity Recognition. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1–15, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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 M. 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.

Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A. Smith, and Jaime G. Carbonell. 2018. Neural cross-lingual named entity recognition with minimal resources. *CoRR*, abs/1808.09861.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online. Association for Computational Linguistics.

Huiyun Yang, Shujian Huang, Xin-Yu Dai, and Jiajun Chen. 2019. Fine-grained Knowledge Fusion for Sequence Labeling Domain Adaptation. 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 4195–4204, Hong Kong, China. Association for Computational Linguistics.

Joey Tianyi Zhou, Hao Zhang, Di Jin, Hongyuan Zhu, Meng Fang, Rick Siow Mong Goh, and Kenneth Kwok. 2019. Dual Adversarial Neural Transfer for Low-Resource Named Entity Recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3461–3471, Florence, Italy. Association for Computational Linguistics.