Robust Named Entity Recognition and Linking on Historical Multilingual Documents

Emanuela Boros, Elvys Linhares Pontes, Luis Adrián Cabrera-Diego, Ahmed Hamdi, Jose G. Moreno, Nicolas Sidère, Antoine Doucet

▶ **To cite this version:**

Emanuela Boros, Elvys Linhares Pontes, Luis Adrián Cabrera-Diego, Ahmed Hamdi, Jose G. Moreno, et al.. Robust Named Entity Recognition and Linking on Historical Multilingual Documents. Conference and Labs of the Evaluation Forum (CLEF 2020), Sep 2020, Thessaloniki, Greece. pp.1-17, 10.5281/zenodo.4068074 . hal-03026969

HAL Id: hal-03026969
https://hal.science/hal-03026969

Submitted on 26 Nov 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

Robust Named Entity Recognition and Linking on Historical Multilingual Documents*

Emanuela Boros[1][0000−0001−6299−9452],
Elvys Linhares Pontes[1][0000−0002−9571−5193],
Luis Adrián Cabrera-Diego[1][0000−0002−9881−9799],
Ahmed Hamdi[1][0000−0002−8964−2135],
Jose G. Moreno[1,2][0000−0002−8852−5797],
Nicolas Sidère[1], and
Antoine Doucet[1][0000−0001−6160−3356]

[1] University of La Rochelle, L3i, F-17000, La Rochelle, France
firstname.lastname@univ-lr.fr
https://www.univ-larochelle.fr
[2] University of Toulouse, IRIT, UMR 5505 CNRS, F-31000, Toulouse, France
firstname.lastname@irit.fr

Abstract. This paper summarizes the participation of the L3i laboratory of the University of La Rochelle in the *Identifying Historical People, Places, and other Entities* (HIPE) evaluation campaign of CLEF 2020. Our participation relies on two neural models, one for named entity recognition and classification (NERC) and another one for entity linking (EL). We carefully pre-processed inputs to mitigate its flaws, notably in terms of segmentation. Our submitted runs cover all languages (English, French, and German) and sub-tasks proposed in the lab: NERC, end-to-end EL, and EL-only. Our submissions obtained top performance in 50 out of the 52 scoreboards proposed by the lab organizers. In further detail, out of 70 runs submitted by 13 participants, our approaches obtained the best score for all metrics in all three languages both for NERC and for end-to-end EL. It also obtained the best score for all metrics in French and German for EL-only.

Keywords: Information Extraction · Named Entity Recognition · Entity Linking

1 Introduction

Identifying historical people, places and other entities is a key task in the automatic understanding of historical newspapers. However, the use of electronic

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0). CLEF 2020, 22-25 September 2020, Thessaloniki, Greece.

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

CEUR-WS.org/Vol-2696/paper_171.pdf

formats for storing text content is relatively new in comparison to the origins of newspapers. For instance, in Europe, the first newspapers appeared at the beginning of the 17^{th} century [25]. Electronic text files started to be widely used since the adoption of operating systems such as MS-DOS in the 1980s [2]. Thus, in the absence of electronic versions of historical newspapers, a common strategy is to recognize the text from digital images of newspapers using optical character recognition (OCR) techniques. In this context, the HIPE 2020 lab at CLEF presented an evaluation campaign with the goal of assessing the recent advances in two major NLP tasks, named entity recognition and classification (NERC) and entity linking (EL), in the context of historical newspapers [5]. This paper presents the participation of the *Laboratoire Informatique, Image et Interaction* (L3i laboratory) at the University of La Rochelle at CLEF HIPE 2020. We developed two new models for NERC and EL. Despite the fact that both models are based on neural networks, there are strong differences between them. Our NERC model is mainly based on the transformer architecture [24] while our EL model is based on a BiLSTM architecture [10]. Our main contributions are three-fold: (1) we propose a pre-processing strategy to mitigate the characteristics of input documents, (2) we extend a transformer-based model for NERC, and (3) we adapt an EL model to a multilingual context. Official results of our participation show the effectiveness of our models over the CLEF HIPE 2020 benchmark.

The remaining of the paper is organized as follows: Section 2 presents the task and the used corpus. Section 3 presents the global architecture of our participation, Section 4.1 presents the pre-processing strategy, while Sections 4 and 5 present individually our NERC and EL systems respectively.

2 HIPE Corpus and HIPE Evaluation

The HIPE corpus [4] 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 dataset was annotated according to the HIPE annotation guidelines [6] which derived from the Quaero[3] annotation guide.

The corpus uses the IOB format with hierarchical information and, provides training, development, and test datasets for each language, except for English. In the case of the latter, the organizers provided only partitions for development and test. In Table 1, we present the statistics regarding the number of named entities found in each dataset. See [3] for a more detailed description of the HIPE dataset.

Regarding the HIPE evaluation, it consists in assessing both tasks, NERC and EL, in terms of Precision (P), Recall (R), and F-measure (F1) at macro and micro levels [14, 3]. Two evaluation scenarios are considered: strict (exact boundary matching) and relaxed (fuzzy boundary matching).

[3] Quaero guidelines: http://www.quaero.org/media/files/bibliographie/quaero-guide-annotation-2011

Table 1: Number of entities for the training, development, and test sets in HIPE 2020 corpora.

Splits	German	English	French
training	3,505	-	6,885
development	1,390	967	1,723
test	1,147	449	1,600

3 L3i NERC-EL Model for Historical Newspapers

In Figure 1, we present the global architecture of our end-to-end NERC-EL model composed of three elements. The first one is a pre-processing module, which reformats the input provided by the organizers. The second element is the NERC module, where we predict the named entities for each language, English, French, and German. The third element is the EL module, where we disambiguate the named entities, and we link them to the Wikidata.

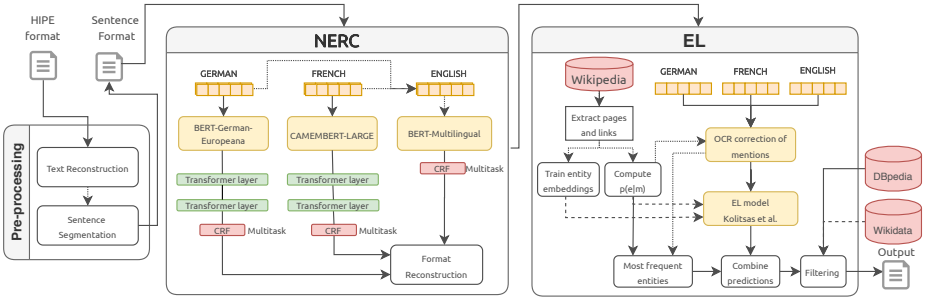


Fig. 1: Global architecture of the NERC and EL proposed models.

In the following sections, we will describe in-depth each of the modules showed in Figure 1.

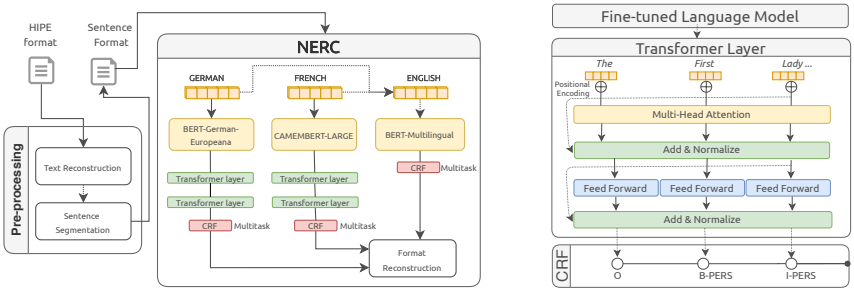
4 Named Entity Recognition and Classification (NERC)

In CLEF HIPE 2020, the NERC task consists in the recognition and classification of entities, such as people and locations, within historical multilingual newspapers. According to the organizers [3], it is composed of two sub-tasks with different levels of difficulty:

– Sub-task 1.1 - NERC coarse-grained: the identification and categorization of entity mentions according to high-level entity types, Person, Location, Organization, Product, and Time.

– Sub-task 1.2 - NERC fine-grained: the recognition and classification of entity mentions at different levels, finer-grained entity types and nested entities, up to one level of depth. It also consists in detecting the components belonging to an entity mention, such as its function, title, honorifics, and name.

Due to the complexity and characteristics of both coarse-grained and fine-grained NERC sub-tasks, we propose the use of a hierarchical, multitask learning approach consisting in a fine-tuned encoder based on *Bidirectional Encoder Representations from Transformers* (BERT) [1]. Our approach includes the use of a stack of Transformer [24] blocks on top of the BERT model for the French and German languages. The multitask prediction layer consists of six separate conditional random field (CRF) layers. The architecture of the model is presented in Figure 2.



(a) NERC architecture for all the languages, including the pre-processing step.

(b) Detailed model proposed for each language.

Fig. 2: The main architecture of the BERT-based model and the additional Transformers (a) is composed of modules stacked on top of each other multiple times. The transformer encoder module (b) mainly consists of multi-head attention and pointwise feed-forward layers.

We decided to use BERT not only because it is easy to fine-tune, but it has also proved to be one of the most performing technologies in multiple NLP tasks [1, 12, 20]. However, while BERT had a major impact in the NLP community, its ability to handle noisy inputs is still an open question [23] or at least requires the addition of complementary methods [16, 19]. More specifically, the built-in tokenizer used by BERT first performs simple white-space tokenization, then applies a Byte Pair Encoding (BPE) based WordPiece tokenization [27]. A 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, the character insertion modification has the minimum influence [23], because the tokenization at the sub-word level of BERT would not change much in some cases, such as "practically" → "practicaally", but the sub-

stitution and deletion errors can hurt the performance of the tokenizer the most due to the generation of uncommon samples, as such as "professionalism" → "pr9fessi9nalism". Thus, these new noisy tokens could influence the performance of BERT-based models[4].

The added layers consist in a stack of Transformer blocks (Transformer encoders). As proposed in [24], this model is a deep learning architecture based on multi-head attention mechanisms with *sinusoidal position embeddings* [5]. 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.

4.1 Data Pre-processing

The HIPE dataset has three different levels of segmentation: article-level, line-level, and newspaper-level. Figure 3 shows an example of the segmentation proposed in the HIPE dataset.

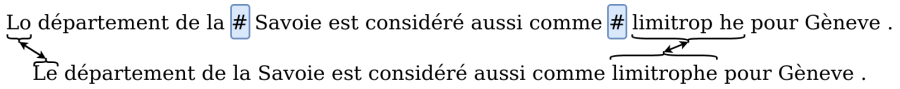


Fig. 3: An example of a French instance from the training data. The upper sentence shows the provided input and the lower sentence contains no OCR errors. "#" represents the segmentation at line-level in historical newspapers. The arrows indicate the matching between the provided sentence and the correct sentence to highlight the OCR limitations.

Since BERT is able to consume only a limited context of tokens (512) and a line-level context would have been too short to grasp, we segment the articles at sentence level. We reconstructed the original text, including hyphenated words, using the miscellaneous annotated column that indicates if a word is split into two or more text lines. Then, the reconstructed text was passed through Freeling 4.1 [18] which determined the boundaries of each sentence.[6]

4.2 Parameters

For the German NERC, we chose as a pre-trained model the `bert-base-german-europeana`. This BERT model was trained using the open-source corpus Euro-

[4] 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.

[5] In our implementation, we used *learned absolute positional embeddings* [8] instead, as suggested by [26]. [24] found that both versions produced nearly identical results.

[6] It should be noted, that the segmentation using Freeling was not flawless. For instance, certain abbreviations were unknown by the tool. Thus, in some cases, Freeling oversegmented the sentences. Nonetheless, these errors were ignored.

peana newspapers[7] [17]. It has been used in other NERC systems for contemporary and historical German texts [22, 21]. Moreover, it has shown an improvement with respect to other NERC systems.

For the French NERC, we relied on a pre-trained CamemBERT [15] model, specifically on the large version, `camembert-large`. Unlike BERT, this French version makes use of a whole-word masking and *SentencePiece* tokenization [11]. Additionally, for `camembert-large`, we found that fine-tuning was sometimes unstable on small datasets, so we ran several random restarts and selected the best model on the development set.

For the English NERC, since no training data was provided, we tackled the task with two approaches. The first one was to train the NERC using the English CoNLL 2003 dataset and the `bert-large-cased` model. The second approach was to use the German and French training data and the pre-trained multilingual BERT model, `bert-base-multilingual-cased`.

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` has L=12, H=768, A=12, `bert-large` 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 [1]. We found that a 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.

4.3 Experiments

The experiments consider two configurations of our previously described model. The first one consists in using only the BERT encoder along with the CRF layers. The second configuration adds the Transformer blocks to the BERT encoder and the CRF layers. In Table 2, we present these experiments per language.

- RUN1: for German, French, and English, the models consist in only the fine-tuning BERT and the CRF layers, with the difference that, for English, we use the CoNLL dataset, and the fine-tuned BERT encoder is the English `bert-large-cased`
- RUN2: for German and French, the models consist in only the fine-tuning BERT, two stacked Transformer blocks, and the CRF layers, while for English, the model is `bert-large-cased`
- RUN3: for English, the model is the one used in RUN1, with the difference that the training data consists of the French and German training data, and the fine-tuned BERT encoder is `bert-base-multilingual-cased`

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

Table 2: The NERC participating COARSE-LIT results for all runs.

Runs	Metrics	German			French			English		
		P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
RUN1	micro-fuzzy	0.838	0.886	0.861	0.909	0.926	0.917	0.775	0.797	0.786
	micro-strict	0.764	0.807	0.785	0.823	0.839	0.831	0.623	0.641	**0.632**
RUN2	micro-fuzzy	0.87	0.886	**0.878**	0.912	0.931	**0.921**	0.774	0.786	0.78
	micro-strict	0.79	0.805	**0.797**	0.831	0.849	**0.84**	0.621	0.63	0.625
RUN3	micro-fuzzy	–	–	–	–	–	–	0.794	0.817	**0.806**
	micro-strict	–	–	–	–	–	–	0.617	0.635	0.626

Table 3: The NERC participating results (all metrics) for the best performing run for each language.

Metrics	German			French			English		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
	COARSE-LIT								
micro-fuzzy	0.87	0.886	0.878	0.912	0.931	0.921	0.794	0.817	0.806
micro-strict	0.79	0.805	0.797	0.831	0.849	0.84	0.617	0.635	0.626
macro_doc-fuzzy	0.879	0.876	0.871	0.933	0.939	0.934	0.782	0.797	0.798
macro_doc-strict	0.782	0.781	0.777	0.852	0.859	0.854	0.635	0.64	0.644
	COARSE-METO								
micro-fuzzy	0.626	0.78	0.694	0.676	0.67	0.673	1.0	0.12	0.214
micro-strict	0.571	0.712	0.634	0.658	0.652	0.655	0.667	0.08	0.143
macro_doc-fuzzy	0.558	0.678	0.686	0.628	0.732	0.718	1.0	0.075	0.533
macro_doc-strict	0.525	0.637	0.645	0.624	0.73	0.715	0.5	0.05	0.333
	FINE-COMP								
micro-fuzzy	0.654	0.768	0.707	0.751	0.827	0.787	0	0	0
micro-strict	0.595	0.698	0.642	0.661	0.728	0.693	0	0	0
macro_doc-fuzzy	0.609	0.719	0.678	0.773	0.833	0.809	0	0	0
macro_doc-strict	0.559	0.649	0.618	0.703	0.757	0.735	0	0	0
	FINE-LIT								
micro-fuzzy	0.734	0.813	0.771	0.843	0.869	0.856	0.733	0.817	0.773
micro-strict	0.629	0.697	0.661	0.772	0.797	0.784	0.547	0.61	0.577
macro_doc-fuzzy	0.754	0.813	0.776	0.871	0.883	0.875	0.742	0.798	0.774
macro_doc-strict	0.644	0.694	0.663	0.799	0.81	0.803	0.584	0.614	0.602
	FINE-METO								
micro-fuzzy	0.659	0.771	0.711	0.626	0.688	0.655	1.0	0.16	0.276
micro-strict	0.601	0.703	0.648	0.618	0.679	0.647	0.75	0.12	0.207
macro_doc-fuzzy	0.595	0.659	0.705	0.558	0.7	0.687	1.0	0.108	0.522
macro_doc-strict	0.562	0.618	0.664	0.556	0.698	0.686	0.667	0.083	0.389
	NESTED								
micro-fuzzy	0.588	0.411	0.484	0.366	0.415	0.389	0	0	0
micro-strict	0.49	0.342	0.403	0.333	0.378	0.354	0	0	0
macro_doc-fuzzy	0.339	0.326	0.413	0.502	0.484	0.521	0	0	0
macro_doc-strict	0.229	0.159	0.252	0.476	0.456	0.491	0	0	0

4.4 Results

From the results in Table 2, we can see the evidence that the BERT-based models with $n\times$Transf achieve, for both German and French languages, higher fuzzy and strict performance values than the stand-alone BERT model.

For a 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). Following this observation, we notice that there is a tendency of performance increase of around 1 percentage F1 points for the $n\times$Transf models (RUN2 for German and French). In Table 3, the highest values for all the coarse and fine metrics are presented.

In the case of English, when comparing RUN1 and RUN2, where the CoNLL 2003 dataset was used for training, with RUN3, where only HIPE German and French datasets were used, we notice that the F1 values are usually degraded by the use of modern datasets in the training process.

In summary, the methods that performed the best for the NERC task were the BERT-based models with n stacked Transformers for German and French. For English, the transfer learning from these two languages was clearly better than the models trained on modern English data.

5 Entity Linking (EL)

Regarding EL, in CLEF HIPE 2020, the task consists in the disambiguation of named entities using two settings:

- End-to-end EL: We do not have prior knowledge of the named entities. Thus, we rely on the information obtained from the NERC system.
- EL-only: We have access to the ground-truth regarding named entities, i.e. types and boundaries.

In both settings, it is necessary to take into account literal and metonymic senses. Furthermore, all the disambiguated named entities have to be linked to the Wikidata knowledge base (KB).

Our EL system is the composition and improvement of two EL approaches (Figure 4). First, we make use of the methodology proposed by [7] to create entity embeddings. Second, we utilize the EL architecture proposed by [10] to disambiguate the candidates. We have modified both EL approaches to support the multilingual aspect of the CLEF HIPE 2020 task.

More precisely, our approach consists of the following four steps which will be elaborated in the subsequent sections:

1. **Building resources**: the setup of a knowledge base per language.
2. **Entity embeddings**: the creation of entity feature representations based on the model proposed by [7].
3. **Entity disambiguation**: the main end-to-end EL model [10].
4. **Candidates filtering**: the post-processing step where several filtering techniques are proposed and studied.

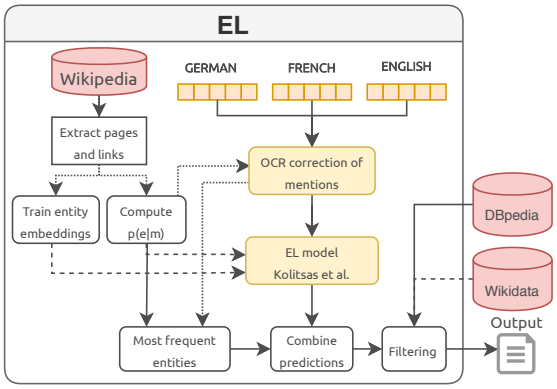


Fig. 4: The proposed model [10] for EL and the post-processing steps.

5.1 Building Resources

We build a KB for English, French, and German, in order to have a richer KB following these steps:

- Retrieve the last language version of the Wikipedia dump.
- Extract titles and ids of Wikipedia pages.
- Extract list of disambiguation pages and redirection pages.
- Calculate the probability entity-map $p(e|m)$ that analyzes how an entity e is related to a mention m based on the number of times that mention refers to that entity.

5.2 Entity Embeddings

We also build a dataset to train entity embeddings for each language, in which case, we use the methodology proposed by [7]. First, we generate two conditional probability distributions per language: the *positive distribution*, which is a probability approximation based on word-entity co-occurrence counts (i.e. which words appear in the context of an entity) and the *negative* one, which was calculated by randomly sampling context windows that were unrelated to a specific entity. Both probability distributions were used for word embeddings alignment with respect to an entity embedding. The *positive distribution* is expected to approach the embeddings of the co-occurring words with the embedding vector of the entity. While the negative probability distribution is used to distance the embeddings of words that are not related to an entity.

5.3 Entity Disambiguation

For the entity disambiguation, our model is based on Kolitsas *et al.*'s work [10], an end-to-end EL model that jointly performs entity linking and entity disambiguation. Besides the simplicity of the model brought by the joint-learning,

the model also takes advantage of the fact that it does not require complex engineered features.

First, for recognizing all entity mentions in a document, Kolitsas *et al.* proposed an empirical probabilistic entity-map[8] $p(e|m)$ to analyze each span m and select top entities e that might be referred by this mention in $p(e|m)$.

The end-to-end EL model starts by encoding every token in the text input by concatenating word and character embeddings that are fed into a Bidirectional Long Short Term Memory (BiLSTM) network. This representation is used to project mentions of this document into a shared dimensional space with the same size as the entity embeddings. These embeddings are fixed continuous entity representations generated separately, namely in the same manner as presented in [7], and aforementioned in Section 5.2.

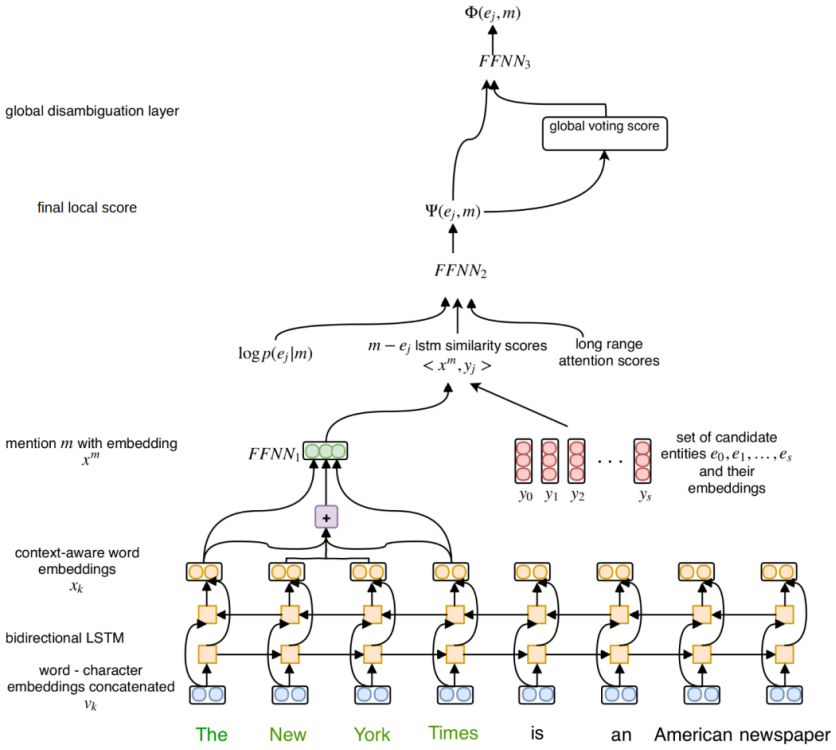


Fig. 5: Global model architecture for the mention "The New York Times". The final score is used for both mention linking and entity disambiguation decisions (Kolitsas *et al.* [10]).

For analyzing long context dependencies of mentions, the authors used the attention model proposed by [7] that produces one context embedding per men-

[8] Calculated from the Wikipedia corpora for each language.

tion based on informative context words that are related to at least one of the candidate entities. Next, the local score for each mention is determined by the combination of the $\log p(e|m)$, the similarity between the analyzed mention and each candidate entity embeddings, and the long-range context attention for the target mention. Finally, a top layer in the neural network promotes the coherence among disambiguated entities inside the same document. Additionally, we provide the five best candidate entities for a mention based on the probability entity-map $p(e|m)$.

5.4 Candidates Filtering

To improve the accuracy of the candidates provided by the EL system, we created a filtering tool based on heuristics and the DBpedia hierarchical structure [13].

Specifically, we used the DBpedia structure to manually specify subsets that represented each named entity type. For instance, the entity type location was associated with categories such as "dbo:Location" and "dbo:Settlement".

These categories were used to determine whether a candidate provided by the EL system had to be positioned at the bottom of the rankings. In other words, candidates that according to DBpedia did not belong to the named entity type were positioned at the bottom of the ranking.[9] For those candidates matching the named entity type[10], we extracted their name in the language of analysis. This name was compared with the entity entry using Fuzzy Wuzzy Weighted Ratio[11]. The most similar candidate to the entity entry was considered to be the most suitable candidate and was positioned at the top.

In the case of person-type entities, we requested to DBpedia their date of birth and extract the year if it was possible.[12] Then, we compared the extracted year of birth with the newspaper publication year, which was provided by the organizers, plus ten years more. If the person entity was born ten years after the publication of the newspaper, we removed completely the candidate.

Furthermore, we created a heuristic that consisted in adding NIL as the last possible candidate. This was done for each named entity unless the EL system proposed candidates with a type different from the named entity one. In this last case, a NIL was inserted between the different types of candidates. For example, if the location "Paris, France" had four candidates entries of type LOC, PERS, LOC, the filter would sort them as LOC, LOC, NIL, PERS. When the EL system proposed only candidates that were different from the named entity type, the filter would position on first place a NIL. These heuristics were based on the idea that if the EL system could not provide a candidate of the same type to the named entity, we might be dealing with an entity without an entry in Wikidata.

For RUN3 in the EL-only task, which will be described in Section 5.5, we proposed as well a filter based on DBpedia along with Wikidata. The reason is

[9] This included candidates that could not be found in DBpedia as well.

[10] In the case the literal and metonymic entities types were discordant, we considered both types as possible.

[11] github.com/seatgeek/fuzzywuzzy

[12] Certain person-type entities, such as music bands, do not have a date of birth.

that the former indexes only a subset of the latter. Thus, to improve the filter, we decided to use Wikidata as a backup knowledge base.

To access DBpedia[13] and Wikidata[14], we utilized their respective SPARQL Endpoint query service.

Table 4: EL results without prior knowledge of mention types and boundaries.

Runs	**Metrics**	**English**			**French**			**German**		
		P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Literal										
RUN1	micro-strict	0.514	0.533	0.523	0.592	0.601	0.597	0.508	0.529	0.518
	micro-relaxed	0.514	0.533	0.523	0.612	0.621	**0.617**	0.53	0.552	0.541
RUN2	micro-strict	0.496	0.506	0.501	0.592	0.602	0.597	0.531	0.538	**0.534**
	micro-relaxed	0.496	0.506	0.501	0.612	0.622	**0.617**	0.553	0.561	**0.557**
RUN3	micro-strict	0.523	0.539	**0.531**	0.594	0.602	**0.598**	0.502	0.528	0.515
	micro-relaxed	0.523	0.539	**0.531**	0.613	0.622	**0.617**	0.524	0.55	0.537
Metonymic										
RUN1	micro-strict	0.172	0.2	**0.185**	0.236	0.402	**0.297**	0.324	0.508	**0.396**
	micro-relaxed	0.172	0.2	**0.185**	0.366	0.625	**0.462**	0.384	0.602	**0.469**
RUN2	micro-strict	0.062	0.04	0.049	0.217	0.339	0.265	0.324	0.508	**0.396**
	micro-relaxed	0.062	0.04	0.049	0.343	0.536	0.418	0.384	0.602	**0.469**
RUN3	micro-strict	0.059	0.04	0.048	0.236	0.402	**0.297**	0.308	0.508	0.383
	micro-relaxed	0.059	0.04	0.048	0.366	0.625	**0.462**	0.364	0.602	0.454

5.5 Experiments

Both entity embeddings and the end-to-end EL method used the pre-trained multilingual MUSE[15] word embeddings of size 300 for all languages in the dataset. We chose the size of 50 for the character embeddings. The German and French models were trained on the HIPE split (Table 1. As the HIPE dataset does not contain training data for English, we trained our English model on the AIDA dataset [9].

In order to overcome or reduce OCR problems, we analyzed several mention variations in order to improve the matching with candidates within the probability entity-map. More precisely, we analyze the following variations: concatenation, lowercase, no punctuation, and the Levenshtein distance between a mention and all candidate mentions within the probability table. In the metonymic sense, the approach used was to annotate the corpus consisted in copying the candidates used for the literal sense.

We implemented three configurations of our EL approach for the EL-only task:

[13] wiki.dbpedia.org/public-sparql-endpoint
[14] query.wikidata.org
[15] https://github.com/facebookresearch/MUSE

– RUN1: for German, French, and English, the output is composed of the candidate entities proposed by [10].
– RUN2: for German, French, and English, the output is composed of the five most frequent candidate entities related to a mention.
– RUN3: for German, French, and English, the output is composed of the candidate entities proposed by [10] and the ten most frequent candidate entities related to a mention. For this run, the filter used not only information from DBpedia but also from Wikidata as indicated in Section 5.4.

We also made three configurations of our end-to-end NERC-EL architecture to recognize and disambiguate entities:

– RUN1: for German, French, and English, the output is composed of entities of NERC RUN1 and the disambiguation method of EL RUN1.
– RUN2: for German, French, and English, the output is composed of entities of NERC RUN2 and the disambiguation method of EL RUN1.
– RUN3: for German and French, the output is composed of entities of NERC RUN1 and the disambiguation method of EL RUN2. For English, the output is composed of entities of NERC RUN3 and the disambiguation method of EL RUN1.

All runs analyze the mention variations and use the filter to select the best five candidate entities among all selected candidate entities by each run.

Table 5: EL results with prior knowledge of mention types and boundaries.

Runs	**Metrics**	**English**			**French**			**German**		
		P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Literal										
RUN1	micro-strict	0.593	0.593	**0.593**	0.64	0.638	**0.639**	0.565	0.564	0.565
	micro-relaxed	0.593	0.593	**0.593**	0.66	0.657	**0.659**	0.588	0.587	0.587
RUN2	micro-strict	0.593	0.593	**0.593**	0.635	0.632	0.633	0.564	0.563	0.564
	micro-relaxed	0.593	0.593	**0.593**	0.654	0.652	0.653	0.587	0.586	0.586
RUN3	micro-strict	0.58	0.58	0.58	0.633	0.63	0.632	0.581	0.582	**0.582**
	micro-relaxed	0.58	0.58	0.58	0.653	0.65	0.652	0.601	0.602	**0.602**
Metonymic										
RUN1	micro-strict	0.286	0.48	**0.358**	0.303	0.446	**0.361**	0.443	0.627	**0.519**
	micro-relaxed	0.286	0.48	**0.358**	0.461	0.679	**0.549**	0.515	0.729	**0.604**
RUN2	micro-strict	0.286	0.48	**0.358**	0.303	0.446	**0.361**	0.443	0.627	**0.519**
	micro-relaxed	0.286	0.48	**0.358**	0.461	0.679	**0.549**	0.515	0.729	**0.604**
RUN3	micro-strict	0.286	0.48	**0.358**	0.297	0.438	0.354	0.431	0.61	0.505
	micro-relaxed	0.286	0.48	**0.358**	0.455	0.67	0.542	0.485	0.686	0.568

5.6 Results

For the EL without prior knowledge of mention types and boundaries, our EL approach depends on the performance of our NERC system to recognize and classify the type of entities in historical documents. The results on all the languages

are presented in Table 4. While RUN1 achieved the best results for metonymic, RUN3 outperformed the other configurations on the literal analysis.

For EL with prior knowledge of mention types and boundaries, our system has access to the ground-truth of NERC entities, i.e. correct span and NERC type for all mentions. Table 5 shows the results. As expected, our EL system achieved better results with the ground-truth information (improvement up to 0.09 and 0.31 in the F1 values for literal and metonymic, respectively). All runs achieved similar results for all languages, with the RUN1 being slightly superior to the other runs for literal and metonymic analysis.

The use of the filter based on DBpedia and Wikidata reduced the performance of the EL system in English and French. This might be due to the increment of noise, such as names of disambiguation pages.[16] Our filter analyses all candidate entities for each mention to order the list of candidates based on their NERC types and names. Since RUN1 and RUN2 provide up to five candidate entities for each mention, these runs are more likely than RUN3 to provide a NIL entry for a mention. For RUN3, the filtering process has a higher probability to find a candidate of the same named entity type as the mention and disambiguates this mention to a less frequent candidate entity in a KB.

6 Conclusions

For the participation of our team (L3i) to the HIPE lab at CLEF 2020, we proposed two neural-based methods for the tasks of NERC and EL. We conclude, for NERC, that the proposed models generally performed well, and that the stacked transformer-based model with a BERT fine-tuned model and additional transformer layers better learned the characteristics of the HIPE historical dataset.

For EL, our neural model combined with the filtering process analyzed the historical mentions and disambiguated them to the Wikidata KB. Combining information from Wikipedia, Wikidata, and DBpedia allowed a thorough analysis of the characteristics of the entities and helped our method to correctly disambiguate mentions in historical documents.

[16] DBpedia does not index disambiguation pages.

Bibliography

[1] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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. pp. 4171–4186 (2019)

[2] Duncan, R.: Advanced MS-DOS Programming. Microsoft Press Redmond, WA (1988)

[3] Ehrmann, M., Romanello, M., Bircher, S., Clematide, S.: Introducing the CLEF 2020 HIPE shared task: Named entity recognition and linking on historical newspapers. In: Jose, J.M., Yilmaz, E., Magalhães, J., Castells, P., Ferro, N., Silva, M.J., Martins, F. (eds.) Advances in information retrieval. pp. 524–532. Springer International Publishing, Cham (2020)

[4] Ehrmann, M., Romanello, M., Clematide, S., Ströbel, P.B., Barman, R.: Language resources for historical newspapers: the impresso collection. In: Proceedings of The 12th Language Resources and Evaluation Conference. pp. 958–968 (2020)

[5] Ehrmann, M., Romanello, M., Flückiger, A., Clematide, S.: Extended Overview of CLEF HIPE 2020: Named Entity Processing on Historical Newspapers. In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) CLEF 2020 Working Notes. Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum. CEUR-WS (2020)

[6] Ehrmann, M., Romanello, M., Flückiger, A., Clematide, S.: Impresso named entity annotation guidelines (version 2.2.0). https://doi.org/10.5281/zenodo.3604227 (2020)

[7] Ganea, O.E., Hofmann, T.: Deep joint entity disambiguation with local neural attention. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 2619–2629 (2017)

[8] Gehring, J., Auli, M., Grangier, D., Yarats, D., Dauphin, Y.N.: Convolutional sequence to sequence learning. arXiv preprint arXiv:1705.03122 (2017)

[9] Hoffart, J., Yosef, M.A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., Weikum, G.: Robust disambiguation of named entities in text. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. pp. 782–792. Edinburgh, Scotland, UK. (2011)

[10] Kolitsas, N., Ganea, O.E., Hofmann, T.: End-to-end neural entity linking. In: Proceedings of the 22nd Conference on Computational Natural Language Learning. pp. 519–529 (2018)

[11] Kudo, T., Richardson, J.: Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. arXiv preprint arXiv:1808.06226 (2018)

[12] Lample, G., Conneau, A.: Cross-lingual language model pretraining. arXiv preprint arXiv:1901.07291 (2019)

[13] Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., Kleef, P.v., Auer, S., Bizer, C.: DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia **6**(2), 167–195. https://doi.org/10.3233/SW-140134
[14] Makhoul, J., Kubala, F., Schwartz, R., Weischedel, R., et al.: Performance measures for information extraction. In: Proceedings of DARPA broadcast news workshop. pp. 249–252. Herndon, VA (1999)
[15] Martin, L., Muller, B., Suárez, P.J.O., Dupont, Y., Romary, L., de la Clergerie, É.V., Seddah, D., Sagot, B.: Camembert: a tasty french language model. arXiv preprint arXiv:1911.03894 (2019)
[16] Muller, B., Sagot, B., Seddah, D.: Enhancing bert for lexical normalization. In: Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019). pp. 297–306 (2019)
[17] Neudecker, C.: An open corpus for named entity recognition in historic newspapers. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). pp. 4348–4352. European Language Resources Association (ELRA), Portorož, Slovenia (May 2016), `https://www.aclweb.org/anthology/L16-1689`
[18] Padró, L., Stanilovsky, E.: FreeLing 3.0: Towards Wider Multilinguality. In: Calzolari, N., Choukri, K., Declerck, T., Doğan, M.U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). pp. 2473–2479. ELRA, Istanbul, Turkey (May 2012)
[19] Pruthi, D., Dhingra, B., Lipton, Z.C.: Combating adversarial misspellings with robust word recognition. In: 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019). pp. 5582–5591. Florence, Italy (2019)
[20] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training (2018)
[21] Riedl, M., Padó, S.: A named entity recognition shootout for german. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 120–125 (2018)
[22] Schweter, S., Baiter, J.: Towards robust named entity recognition for historic german. arXiv preprint arXiv:1906.07592 (2019)
[23] Sun, L., Hashimoto, K., Yin, W., Asai, A., Li, J., Yu, P., Xiong, C.: Advbert: Bert is not robust on misspellings! generating nature adversarial samples on bert. arXiv preprint arXiv:2003.04985 (2020)
[24] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
[25] Weber, J.: Strassburg, 1605: The origins of the newspaper in europe. German history **24**(3), 387–412 (2006)
[26] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Brew, J.: Huggingface's transformers: State-of-the-art natural language processing. ArXiv **abs/1910.03771** (2019)

[27] Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al.: Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144 (2016)