Project Number: **770299**

NewsEye:

A Digital Investigator for Historical Newspapers

Research and Innovation Action
Call H2020-SC-CULT-COOP-2016-2017

D3.5: Named Entity Recognition and Linking (final)

Due date of deliverable: M24 (30 April 2020)

Actual submission date: 28 April 2020

Start date of project: 1 May 2018 **Duration:** 36 months

Partner organization name in charge of deliverable: ULR

Project co-funded by the European Commission within Horizon 2020		
Dissemination Level		
PU	Public	PU
PP	Restricted to other programme participants (including the Commission Services)	-
RE	Restricted to a group specified by the Consortium (including the Commission Services)	-
CO	Confidential, only for members of the Consortium (including the Commission Services)	-

Revision History

Document administrative information		
Project acronym:	NewsEye	
Project number:	770299	
Deliverable number:	D3.5	
Deliverable full title:	Named Entity Recognition and Linking (final)	
Deliverable short title:	Named Entity Recognition and Linking (final)	
Document identifier:	NewsEye-T31-D35-NE_Recognition_and_Linking-Submitted-v3.0	
Lead partner short name:	ULR	
Report version:	V3.0	
Report preparation date:	28.04.2020	
Dissemination level:	PU	
Nature:	Report	
Lead author:	Ahmed Hamdi (ULR) and Elvys Linhares Pontes (ULR)	
Co-authors:	Antoine Doucet (ULR)	
Internal reviewers:	Eva Pfanzelter (UIBK-DEA), Lidia Pivovarova (UH-CS)	
Status:		Draft
		Final
	x	Submitted

The NewsEye Consortium partner responsible for this deliverable has addressed all comments received, making changes as necessary. Changes to this document are detailed in the change log table below.

Change Log

Date	Version	Editor	Summary of changes made
23/03/2020	0.1	Ahmed Hamdi and Elvys Linhares Pontes (ULR)	First draft
02/04/2020	0.2	Ahmed Hamdi and Elvys Linhares Pontes (ULR)	Minor improvements following ULR-internal feedback
03/04/2020	0.3	Antoine Doucet (ULR)	Suggestions and text improvements
10/04/2020	1.0	Ahmed Hamdi, Elvys Linhares Pontes and Antoine Doucet (ULR)	Final draft, made available to internal reviewers
24/04/2020	2.0	Ahmed Hamdi, Elvys Linhares Pontes and Antoine Doucet (ULR)	Final version, with reviews taken into account
28/04/2020	3.0	Antoine Doucet (ULR)	Final adjustments towards submission

Executive summary

The overall objective of WP3 is the semantic text enrichment of individual documents and their contents. This deliverable is the final report on Task T3.1, which is concerned with the recognition and linking of named entities (NEs), predefined real-world objects to be recognised from text written in natural language. Task T3.1 addresses two subtasks: named entity recognition (NER) and named entity linking (NEL).

A first deliverable at M12 (D3.2) focused on the state of the art and impact of OCR on the performance of NER and NEL. The present report is the final presentation of our work on Task T3.1, describing robust to noise and language-independent approaches.

The output of Task T3.1 will be used in many ways. First the semantic enrichment will continue with the detection of stance towards extracted named entities in Task T3.2, and the detection of events in Task T3.3. This output from WP3 will result in a cross-lingual knowledge base that will be accessible directly to users through the demonstrator, and that will feed the analysis tools of WP4 and the personal research assistant (WP5).

This report is organised in two main parts, corresponding to our work on named entity recognition and on named entity linking. In both cases, we evaluate and compare the state of the art to our approaches over historical data. Compared to the baseline, our NER approach achieved relative improvements of 21%, 9%, 31% and 48% on the Finnish, French, German and Swedish data, respectively. Our NEL approach provided an improvement over the baseline on the French and German CLEF-HIPE datasets (20.9% and 3.9%, respectively) and on the Finnish, French, German, and Swedish NewsEye datasets (867.7%, 79.2%, 101.7%, and 12.6%, respectively).

Contents

Named Entity Recognition and Linking

This work is concerned with the recognition and linking of NEs from digitised historical newspapers published between 1850 and 1950. Named entities are among the most relevant information that can help to properly index digital documents and easily retrieve them. However, most digitised documents are indexed through a noisy version produced by an optical character recognition (OCR) system. The noisy version contains numerous OCR errors that change the content of these documents and naturally make their access more difficult in digital libraries.

Unlike contemporary data that have a large number of NER and NEL resources and tools, historical documents face the problem of lacking annotated resources. Contemporary resources are not suitable to build accurate tools over historical data because of variations in orthographic and grammatical rules, not to mention the fact that the names of persons, organisations and places are significantly changing over time.

In order to analyse the problems resulting from OCR and to assess the challenges related to the processing of historical data, we first test state-of-the-art NER techniques over several historical datasets and then propose effective techniques that remedy state-of-the-art techniques and subsequently allows achieving better performance with historical datasets; this work is presented in Section 1. Following the same idea about the challenges of processing historical data, we extend the analysis of NER in these documents by linking these entities to a knowledge base. We tested a state-of-the-art system to disambiguate named entities to a knowledge base. Then, we compared this system to our approach for evaluating the performance of NEL systems on historical newspapers; this work is presented in Section 2 of this deliverable. We conclude this report in Section 3, followed by Named Entity and Stance Annotation Guidelines presented in Appendix A.

1. Named Entity Recognition

Named entity recognition (NER) is a natural language processing (NLP) task that aims to locate important names and proper names in a given text and to categorise them into a set of predefined classes. Typical NER tag sets define three classes for named entity labelling: persons, locations and organisations [1]. In the NewsEye project, additionally to these classes, NER targets a class including human products and specifies a subtype for the class person when it corresponds to the author of an article. In the context of newspapers, it is indeed very useful to be able to differentiate the person(s) mentioned in an article from the person(s) who wrote and signed the article.

In NewsEye, the NER task is focused on the extraction of named entities from newspaper articles published between 1850 and 1950. There are three key challenges that needed to be addressed: first, texts were produced using automated optical character recognition (OCR) technology which tends to produce a rather high degree of errors in the recognition of words especially historical ones. The OCR quality impacts the effectiveness of NER systems mostly when the OCR error rates are relatively high [2]. Second, several spelling variations can appear in historical texts compared to contemporary datasets. Third, the lack of annotated resources from historic origin does not allow achieving competitive results compared to contemporary results.

In a preliminary report on the task of named entity recognition and linking (Deliverable D3.2), we tested the impact of OCR noise over four state-of-the-art NER systems with the objective to measure the impact

of OCR quality on their performance. By synthesising different levels and types document degradation, we were able to determine that neural network methods are the most robust to OCR noise, something we confirmed with the actual historical newspaper dataset provided and annotated by the national library of Finland [3].

However, neural networks require large resources to reach good results. The existing resources based on historical data are unfortunately few and small, unlike contemporary datasets, which are large and rich-resourced. For this reason, we will first test the existing NER accurate systems based on results and findings from the deliverable D3.2 over historical data. This will define our baseline. Then, we will take advantage of a large amount of available contemporary datasets and use transfer learning techniques to improve the baseline results.

This part of the deliverable is organised as follows: Section 1.1 introduces NER approaches on named entity recognition especially those dealing with historical data. Section 1.2, gives an overview of the baseline systems (cf. Section 1.2.1) and describes the transfer learning technique (cf. Section 1.2.2). In Section 1.3 we present the datasets used for training and testing both methods then we compare and discuss the results in Section 1.4 .

1.1. An overview of named entity recognition

NER systems aim to assign a sequence of labels for a given sequence of words. Each word is a token in a sequence to be assigned a label (e.g., PER for persons, LOC for locations, ORG for organisations and O for words that are not named entities). The sentence "*John lives in New York*", for instance, has to be labeled as follows: "*PER O O LOC LOC*". In this section, we first summarise the main NER approaches, then we describe the most important NER works dealing with historical data.

1.1.1. Named entity recognition approaches

The first NER system has been proposed in the message understanding conference (MUC) in the 1990's [4], and early approaches relied on rule-based approaches. Rules used in those systems are defined by humans and based on dictionaries, linguistic descriptors and trigger words. The word "*Mr*" for example generally triggers a named entity of type person.

While rule-based techniques do not require annotated resources to define rules, they need huge manual efforts and a lot of time and human expertise to be extracted and handled. Rule-based approaches cannot, therefore, be easily adapted to new types of texts or entities. To overcome this problem, efforts on NER are now largely dominated by machine learning techniques such as fully supervised learning, semi-supervised learning, unsupervised learning.

Fully supervised approaches to NER include support vector machines (SVM) [5], maximum entropy models [6], Decision Trees [7] as well as sequential tagging methods such as Hidden Markov Models [8], and Conditional Random Fields (CRFs) [9, 10, 11, 12]. These approaches similarly to rule-based methods rely on handcrafted features, which are challenging and time-consuming to develop and may be costly to update and generalise to new data.

More recently, neural networks have been shown to outperform other supervised algorithms for NER.

The first deep neural network-based learning system has been developed in 2011 [13]. It reached very competitive results for NER in comparison to previous machine learning works. Therefore, many NER systems using neural networks have been proposed and have shown their abilities to outperform all previous systems [14, 15, 16]. The effectiveness of NER systems using neural networks is due to their ability to be adapted and generalised. These systems can jointly learn effective features with model parameters directly from the training dataset, instead of relying on handcrafted features developed for a specific dataset. Several related works showed that word embedding techniques impact the effectiveness of deep-learning systems on named entity recognition [17, 18].

Among the variety of neural network architectures applied for NER, many works have used a bidirectional long short term memory (BLSTM) and achieved very good results [19, 20]. BLSTM methods have also shown their effectiveness to handle the NER task when combined with a top-level CRF layer [21, 22, 23]. In this work we have, therefore, chosen to use BLSTM-CNN-CRF [24] in order to define our baseline (cf. Section 1.2.1). This system outperforms the other BLSTM NER systems tested and reported in Deliverable D3.2.

1.1.2. Named entity recognition from historical data

Most of NER systems have been proposed to process contemporary and clean data. Few studies have been devoted to extracting named entities from historical data [25, 26].

Rodriquez *et al.* [27] reported that manual correction of OCR output does not have a very observable improvement on NER results. Other studies interested to named entity extraction from digitised historical journals [28], broadcast news [29] and religious monologues, scientific books and medical emails [30]. In [31], authors presented a complete framework for named entity recognition for both contemporary and historical German using transfer learning technique. They used a combination of BLSTM (that obtain good performances when data quality and quantity are sufficient, such as contemporary datasets) with a CRF as a top layer to achieve state-of-the-art performance for historical datasets with fewer samples that contain noise.

Dealing with noisy data, several efforts have been devoted to extracting named entities from diverse text types such as outputs of automatic speech recognition (ASR) systems [32, 33], informal SMS and noisy social network posts [34]. Palmer and Ostendorf [35] for example described an approach for improving named entity extraction from ASR systems outputs by explicitly modelling errors through the use of confidence scores. In a similar setting, Miller *et al.* [36] have studied the performance of named entity extraction under a variety of spoken and OCRed data. They trained the IdentiFinder system [37] on both clean and noisy input material, performance degraded linearly as a function of word error rates. They concluded that results may lose about eight points of F-score with only 15% of word error rate.

In this work, we follow a similar idea as Riedl *et al.* [31]. We take advantage of the availability of large NER contemporary corpora to train initially NER models and then adapt them using transfer learning for processing historical data. However, unlike them, our study targets more languages and datasets.

1.2. Methodology

The development of effective NER tools require the availability of sufficient training data [38]. However this requirement is not always satisfied especially with new types of text such as historical data or domain specific. To face the problem of insufficient training data, two solutions are available. The first one is rather obvious: to create the missing training data in large amounts. The second solution is to rely on existing resources that are sufficiently related to the problem at hand, and to take advantage of that relatedness to learn adequate knowledge.

In this work, we explore the two options. In collaboration with the NewsEye partners, a NER ground truth based on the NewsEye collections is being created. We also take advantage from the CLEF-HIPE NER resource which is close-related to the NewsEye data, developed and made publicly available in 2020. Our baseline consists of training and testing NER systems on these datasets. We then investigate the possibility of exploiting NER contemporary corpora and transfer learning to reach better results over the NewsEye and the CLEF-HIPE datasets.

1.2.1. Baseline systems

As mentioned in Section 1.1, BLSTM models demonstrate the ability to effectively handle sequence labelling tasks, particularly named entity recognition. As mentioned in the introduction of Section 1, our earlier benchmarking of state-of-the-art NER methods over noisy OCRed text showed that neural network approaches were the most adequate. BLSTM NER systems were particularly robust to noise when processing OCRed inputs, especially with a CRF top layer. For these reasons, we use in this work the BLSTM-CNN-CRF system [24].

This NER system converts the input sequence of words into a sequence of fixed-size vectors $(x_1,x_2,...,x_n)$ and returns another sequence of vectors $(h_1,h_2,...,h_n)$ that represents named entity labels at every step of the input. Long Short-Term Memory networks [39, 40] compute a representation of the context of each input word. The model uses a forward LSTM that represents the left context and a backward LSTM encoding the right context. The forward and backward LSTM pair is referred to a bidirectional LSTM. A CRF layer (cf. Figure 1) finally allows generating the most probable sequence of predicted labels from surrounding words.

BLSTM-CNN-CRF introduces character-level features using a convolutional neural network (CNN) engine (see Figure 1). This system adds to each word vector a new character-based feature vector. In order to extract the character feature vectors, the model employs a convolution and a max-pooling layer. The LSTM networks encode then the concatenation of word vectors and their corresponding character vectors CNN outputs. Finally, the output vectors of LSTM are decoded into the best label sequence using the CRF top layer.

We used the FastText[1] pre-trained word embedding models that are available for 157 languages [41]. While the word embeddings are pre-trained, the character embeddings are trained at the same time as the training of the model. To remedy issues with out-of-vocabulary (OOVs) words, we use both character- and subword-based word embeddings computed with FastText [42]. This method is able to retrieve embeddings for unknown words by incorporating subword information.

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

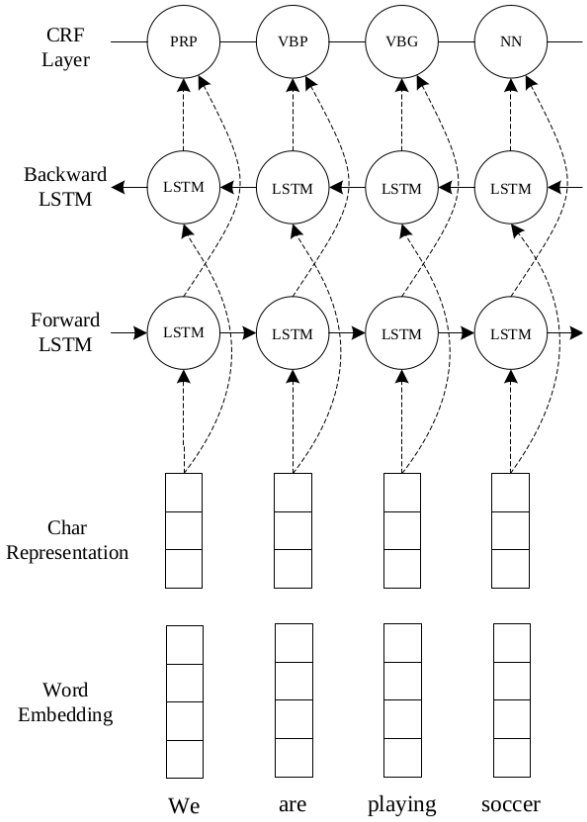


Figure 1: Main architecture of the BLSTM-CNN-CRF. The character representation vector is concatenated with the word embedding before being fed into the BLSTM network. Dashed arrows indicate the dropout layers applied on both the input and output vectors of BLSTM (Ma *et al.* [24]).

1.2.2. Transfer learning technique

Transfer learning has been studied for a long time. However, there is no standard definition of transfer learning in the literature [43]. We follow the definition from [44]: transfer learning aims at performing a task on a target dataset using some knowledge learned from a source dataset. More precisely, a model in a specific task can be trained on one corpus from a source domain and at some point, it switches to another corpus from the target domain on which the task is evaluated. The idea has been applied in many fields such as speech recognition [45], biomedical [46] and finance [47].

For the NER task, large available corpora are almost contemporary while the historical data are small and rare. To process historical data, transfer learning can, therefore, be a good solution. In our scenario for the NewsEye project, we start by training on large contemporary “source” corpora until convergence and then train additional epochs on the NewsEye “target” corpora.

As a starting point, we trained three models using three corpora fr-WikiNER [48], de-GermEval [49] and fi-FiNER [50] for French, German and Finnish, respectively. As shown in Figure 2, each of these source models share all the parameters and feature representations of the neural networks with the target models, including the word and character embeddings, the word-level layer, the character-level layer and the CRF layer.

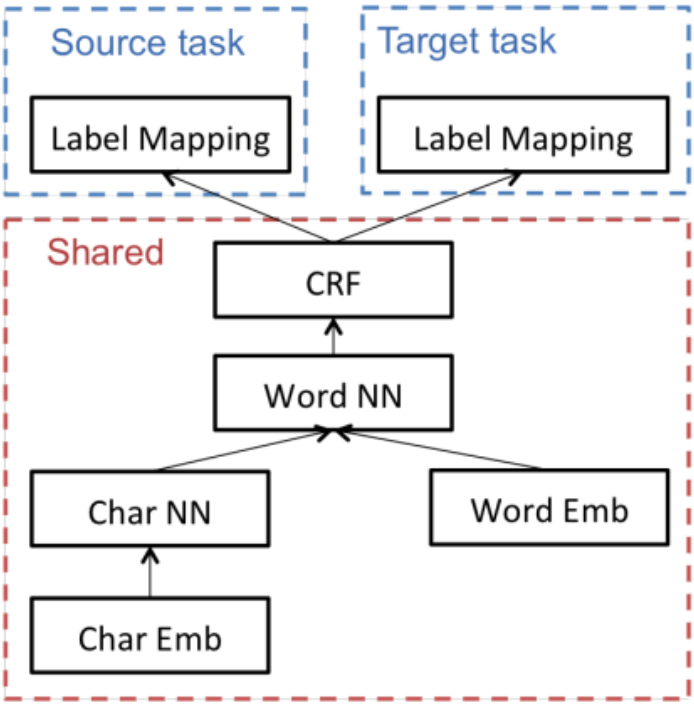


Figure 2: Transfer model used for cross-domain transfer where label mapping is possible (Yang *et al.* [51]).

1.3. Datasets overview

As part of the NewsEye project, we aim to extract named entities from historical newspaper articles in four languages: French, German, Finnish and Swedish. In order to assess our work, we use three collections of data: the NewsEye produced datasets (Section 1.3.1), external historical datasets (Section 1.3.2) and large contemporary datasets (Section 1.3.3).

1.3.1. NewsEye data

To address the lack of dataset that are perfectly suited for the needs of the NewsEye project, notably datasets with both OCR and NER groundtruth matching the needs of NewsEye users, we launched the creation of NewsEye datasets based on the NewsEye collections and languages.

Groundtruth creation. An internal working group was created in 2019 to define the NE categories that would match the needs of the different types of NewsEye users. At the same time, we were developing synergies with the Swiss-Luxembourg Impresso project, also focused on historical newspapers, and in which guidelines were also defined for the creation of annotated datasets of NER and NEL in French and German (this eventually led to the CLEF-HIPE datasets, described in Section 1.3.2). Having annotations compatible across projects would be beneficial for the community at large, in particular for both projects since datasets produced in one project could be used in the other.

We therefore built the annotation guidelines in a concerted manner, and the NewsEye NE annotation guidelines actually started out as a branch of the Impresso NE annotation guidelines. The resulting

NewsEye NE annotation guidelines are provided in Appendix A. Apart for a few fine-grained variations, the main difference with Impresso guidelines is that NewsEye guidelines focus on NE main types and ignore most of the subtypes defined in the Impresso guidelines. The only exception is the subtype *pers.articleauthor* which is kept to recognise authors of newspaper articles, as explained earlier.

Four main types and one subtype of named entities are defined in NewsEye:

- person (PER): individual or group of persons;
  - authors of articles (PER.articleauthor) which indicate authors' names or initials.
- location (LOC): address, territory with a geopolitical border such as city, country, region, continent, nation, state or province;
- organisation (ORG): commercial, educational, entertainment, government, media, medical-science, non-governmental, religious, sports;
- Human production (PROD): we only focus on media products such as newspapers, magazines, broadcasts, etc.

Analysis of NewsEye datasets. Once the guidelines were compiled, as part of Task T1.3 on data generation, partners in UIBK-DEA adapted the Transkribus tool to allow for NE annotations and prepared datasets to be annotated, following up on technicalities. ULR took care of answering numerous questions of annotators on a dedicated Slack channel and correspondingly adjusting guidelines.

Table 1 summarises the NewsEye NE dataset v1, showing the distribution of named entities according to their types.

			named entities				
language	corpus	tokens	total	PER	LOC	ORG	PROD
German	de-NewsEye	168,253	8,845	2,414	3,987	2,405	39
Finnish	fi-NewsEye	48,502	2,669	1,057	1,166	332	114
French	fr-NewsEye	241,071	10,327	4,700	4,046	1,323	258
Swedish	sv-NewsEye	49,595	2,444	996	1,147	188	113

Table 1: Statistical description of the NewsEye NE dataset v1

In order to evaluate the inter-annotator agreement in the NewsEye datasets, several pages from each corpus have been annotated twice by two groups of native speakers of the concerned language. We then compute the IAA using the Kappa coefficient introduced by Cohen [52]. Table 2 shows the inter-annotator agreement in the NewsEye datasets and describes the distribution of annotations between the two groups. For each NE type annotated by one group, we indicate how it was annotated by the other group.

Table 2 shows very satisfactory annotator agreement, with IAA between 0.83 and 0.93 depending on the language dataset. This is also shown with higher numbers in the diagonal cells for persons, locations and a bit less for organisations. In few cases, named entities are associated with two different types by the two groups. This indicates that guidelines distinguish well the different types of NEs. The annotation process triggered many questions from annotators, which created a virtuous circle or clarification of the guidelines, defining rules for ambiguous cases and contributing to improve the consistency of the annotations, and thus the quality and the usefulness of the dataset.

de-NewsEye IAA = 0.91		2^{nd} group				
		PER	LOC	ORG	PROD	Total
1^{st} group	PER	**85**	3	1	0	89
	LOC	2	**279**	8	0	289
	ORG	3	9	**106**	0	118
	PROD	0	0	0	**5**	5
	Total	90	291	115	5	501

fi-NewsEye IAA = 0.93		2^{nd} group				
		PER	LOC	ORG	PROD	Total
1^{st} group	PER	**212**	0	1	0	213
	LOC	2	**15**	9	0	26
	ORG	0	0	**98**	0	98
	PROD	0	0	0	0	0
	Total	214	15	108	0	337

fr-NewsEye IAA = 0.90		2^{nd} group				
		PER	LOC	ORG	PROD	Total
1^{st} group	PER	**303**	0	0	0	303
	LOC	2	**82**	12	0	96
	ORG	6	0	**33**	0	39
	PROD	0	0	1	**7**	8
	Total	311	82	46	7	446

sv-NewsEye IAA = 0.83		2^{nd} group				
		PER	LOC	ORG	PROD	Total
1^{st} group	PER	**126**	1	4	0	131
	LOC	0	**15**	2	0	17
	ORG	1	2	**7**	0	10
	PROD	0	0	0	**5**	5
	Total	127	18	13	5	163

Table 2: Distribution of annotations according to the NE types in the NewsEye dataset v1

1.3.2. External historical data

Few NER resources have been built on historical data for French, German or Finnish. They are described below.

1. **CLEF-HIPE corpora**: three corpora are proposed for the CLEF-HIPE 2020 shared task on named entity recognition and linking[2], and produced as a result of the Swiss-Luxembourg Impresso project[3]. The corpora are extracted from newspaper articles of the last two centuries in three languages: English, French, and German. In this work, we use the French and the German corpora. As we mentioned above, the main advantage of these data is that they follow similar guidelines as those of the NewsEye project. We believe that these annotated corpora are the most similar to NewsEye data and will produce closely related results. The French corpus consists of 186,696 tokens while the German corpus contains 123,137 tokens. The number of named entity mentions is 7,458 and 4,704 in the French and German corpora respectively.
2. **NLF corpus**: the corpus is provided by the National Library of Finland (NLF). It consists of Finnish historical newspapers and journal collections from the period 1771–1929 [3]. The corpus contains around 450K tokens, among which more than 30K are named entities manually annotated. The corpus defines two classes to categorise named entities: PER for names of persons and LOC for locations.

To maintain consistency with the NewsEye types of NEs, we standardised the tagset of all the corpora to the NewsEye four-category set (PERS, LOC, ORG, and PROD). All the other named entity classes are ignored. Table 3 statistically summarises all the external historical datasets used in this work.

[2] https://github.com/impresso/CLEF-HIPE-2020
[3] https://impresso-project.ch/

language	corpus	tokens	named entities				
			total	PER	LOC	ORG	PROD
Geman	de-CLEF-HIPE	123,137	4,704	1,598	2,411	530	165
French	fr-CLEP-HIPE	186,696	7,458	2,955	3,420	846	237
Finnish	fi-NLF	397,227	18,233	7,801	10,431	–	–

Table 3: Statistical description of external historical NER corpora

1.3.3. Contemporary corpora

As described in Section 1.2, besides historical corpora, NER corpora based on contemporary data are required to build initial models for transfer learning. In this work, we selected for each language of the NewsEye project one contemporary corpus.

1. The fr-WikiNER corpus[4] is extracted from Wikipedia's articles. It contains about 500K tokens among them 31,070 are named entities.
2. The de-GermEval corpus[5] sampled data from German Wikipedia and News Corpora as a collection of citations. The dataset covers over 31,000 sentences corresponding to over 590k tokens among them around 33k are named entities.
3. The fi-FiNER corpus[6] is collected from news articles with a manually prepared named entity annotation. The text material was extracted from the archives of Digitoday[7], a Finnish online technology news source. The corpus consists of 953 articles which cover 204,094 word tokens among them 16,180 are named entities.
4. The sv-WebNews corpus[8] is collected from Swedish Gazetters. It is a semi-automatically annotated corpus. Annotations have been predicted by CoreNLP[9] [53] and then manually corrected and reviewed by two Swedish native speakers. The corpus contains about 8,000 sentences, 155,333 tokens and 5,184 named entities.

Each contemporary dataset defines a NER tagset composed of 4 labels: person, location, organisation and miscellaneous[10]. As the external historical datasets, we standardised the tagset of all the contemporary corpora to the NewsEye four-category set (PERS, LOC, ORG, and PROD). Table 4 summarises all the datasets used in this work. We show the number of NE mentions and their distribution according to their classes, for each dataset.

[4] https://figshare.com/articles/Learning_multilingual_named_entity_recognition_from_Wikipedia/5462500
[5] https://sites.google.com/site/germeval2014ner/data
[6] https://github.com/mpsilfve/finer-data
[7] https://www.digitoday.fr
[8] https://github.com/klintan/swedish-ner-corpus
[9] https://stanfordnlp.github.io/CoreNLP/
[10] This type includes all NEs not belonging to the other three types

language	corpus	tokens	named entities				
			total	PER	LOC	ORG	PROD
Geman	de-GermEval	590,984	33,397	10,348	15,028	8,021	–
French	fr-WikiNER	500,231	31,070	9,244	17,632	4,194	–
Finnish	fi-FiNER	240,094	16,180	2,622	2,539	11,019	–
Swedish	sw-FiNER	155,333	5,184	2,199	1,791	1,194	–

Table 4: Statistical description of NER contemporary corpora

1.4. Experiment and results

In order to asses our work we used traditional metrics (Precision, Recall and F1-score) to evaluate NER systems. Precision P is the rate of named entities correctly recognised by the system. Recall R is the rate of named entities present in the corpus that are found by the system. An extracted named entity is considered correct only if it is an exact match of the corresponding entity in the test corpus. The F1-score is the harmonic mean between precision and recall:

$$F1 = \frac{2 * P * R}{P + R} \tag{1}$$

As described in Section 1.3, for each language (French, Finnish, and German) we have at least two datasets, one contemporary and another historical. Each dataset is divided into three parts: 80% of the data for training and each 10% for development and testing.

1.4.1. Baseline method

Our first evaluation consists of running the baseline systems over historical datasets. We performed therefore for each language a cross-corpus evaluation in order to show the F1-score of NER systems over historical data when we trained models on contemporary datasets and also when we trained them over historical datasets. The best F1-scores for each dataset are our baselines. We mark them in bold.

Table 5 shows that the best results are achieved when testing on the same dataset used for training. The NER system clearly showed some limits on processing historical data when trained on small training datasets of the same nature and even when they are trained on large contemporary datasets. Best results are achieved when NER models are trained and tested on data of the same collection. Models built on contemporary datasets are clearly not suitable for processing historical data. They almost give the lowest results when we test on the NewsEye dataset or on the CLEF-HIPE dataset.

1.4.2. Transfer learning

The second evaluation consists of using transfer learning techniques for first training models on contemporary datasets and then to adapting it on historical datasets. We train models on the source corpora until convergence, and then we train few additional epochs on the “target” corpus from the domain on

Language	Train	Test					
		de-NewsEye			de-CLEF-HIPE		
		P	R	F1	P	R	F1
German	de-GermEval	40.89	32.75	36.37	59.08	32.05	41.56
	de-NewsEye	53.37	36.19	**43.13**	47.44	24.50	32.31
	de-CLEF-HIPE	56.28	32.00	40.80	58.16	32.72	**41.88**
		fr-NewsEye			fr-CLEF-HIPE		
		P	R	F1	P	R	F1
French	fr-WikiNER	24.58	30.15	27.08	34.87	51.49	41.58
	fr-NewsEye	64.64	47.04	**54.45**	52.68	45.34	48.74
	fr-CLEF-HIPE	41.83	33.53	37.22	74.21	74.80	**74.51**
		fi-NewsEye			fi-NLF		
		P	R	F1	P	R	F1
Finnish	fi-FiNER	27.69	18.41	22.12	46.69	38.85	42.41
	fi-NewsEye	35.96	25.80	**30.04**	40.98	18.43	25.42
	fi-NLF	39.22	22.36	28.48	87.02	82.03	**84.45**
		sv-NewsEye					
		P		R		F1	
Swedish	sv-WebNews	34.44		29.81		31.96	
	sv-NewsEye	55.22		28.67		**37.74**	

Table 5: Cross-Corpus NER Performance using the baseline system

which we evaluate. One advantage of transfer learning is that models do not require a lot of time to converge. Figure 3 shows that few epochs are sufficient to reach the best NER results.

In our scenario, we start by training on a large contemporary corpora until convergence and then train few additional epochs on the historical corpus from the domain on which we evaluate.

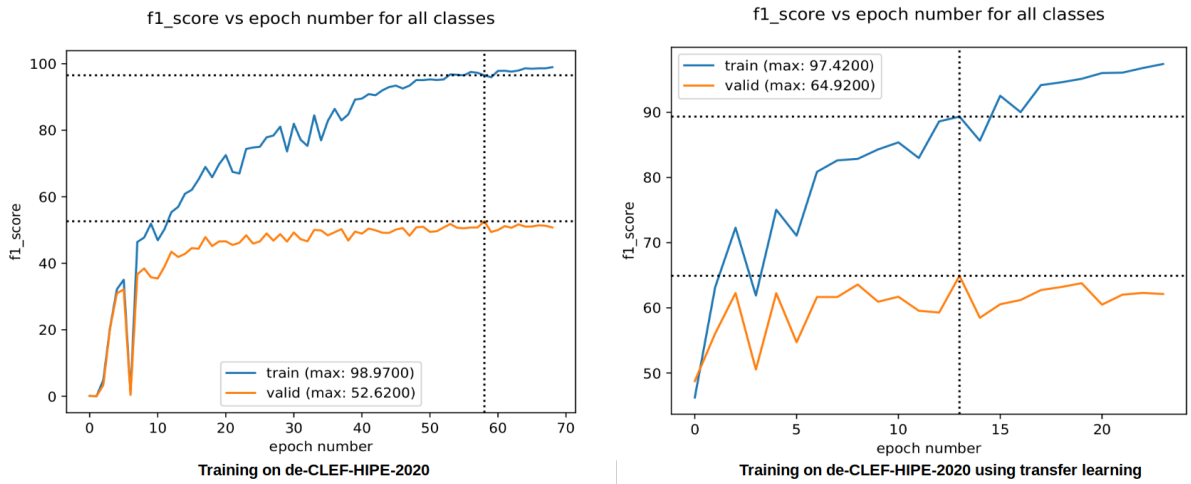


Figure 3: Number of epochs required for training with and without transfer learning

	Train	Transfer	Test	
German			de-NewsEye	de-CLEF-HIPE
	de-GermEval	de-NewsEye	53.74	–
		de-CLEF-HIPE	–	56.58
French			fr-NewsEye	fr-CLEF-HIPE
	fr-WikiNER	fr-NewsEye	57.13	–
		fr-CLEF-HIPE	–	77.30
Finnish			fi-NewsEye	fi-NLF
	fi-WikiNER	fi-NewsEye	36.39	–
		fi-NLF	–	85.43
Swedish			sv-NewsEye	
	fi-wikiNER	sv-NewsEye	55.98	

Table 6: NER F1-score using transfer learning

The results in Table 6 show significant improvements for the NewsEye datasets as well as the CLEF-HIPE datasets. Combining contemporary sources with historic target corpora yields consistent benefits. NER F1-scores on the NewsEye datasets improved across all the languages, on the Finnish dataset for example the NER F1 score increases from 30.04% to 36.39% while on the Swedish data the NER F1-score jumps to 55.98% while it was 37.74% using the baseline system. Over the CLEF-HIPE, NER F1-scores have also been increasing from 41.88% to 56.71% on German and from 74.51% to 77.30% on French. Results on fi-NLF showed a minor improvement, presumably because the data are sufficiently large to build robust NER systems. We conclude that transfer learning is beneficial for NER on historical data, especially when training data for the target domain are small.

1.4.3. Multiple transfer learning

We conducted a third experiment using multiple transfer learning for German and French. It consists of building initial models on contemporary corpora, then we use two consecutive transfer learning on historical data where the last transfer is made on data from the domain of the test data. Regardless of the language, the multiple transfer learning allowed us to improve the baseline results over all the historical datasets (see Table 7). On German data, for example the F1-score jumps from 43.13% to 56.71% on the NewsEye dataset and from 41.88% to 62.97% on the CLEF-HIPE dataset which represent a relative improvement of 31% and 50% respectively over the baseline. For the French data, the F1-score increases from 54.45% to 59.28% on the NewsEye data and from 74.51% to 80.97% with the CLEF-HIPE data, a relative improvement of 9% on both datasets.

	Train	1^{st} transfer	2^{nd} transfer	Test	
German	de-GermEval	de-NewsEye	de-CLEF-HIPE	de-NewsEye	de-CLEF-HIPE
		de-CLEF-HIPE	de-NewsEye	–	62.97
				56.71	–
French	fr-Wikiner	fr-NewsEye	fr-CLEF-HIPE	fr-NewsEye	fr-CLEF-HIPE
		fr-CLEF-HIPE	fr-NewsEye	–	80.97
				59.28	–

Table 7: NER Performance using multiple transfer learning

The transfer learning also improved the accuracy of predicting all the NE classes. The lowest F1 scores are achieved for the label organisation (cf. Table 8). We obtain an F1 score for this label of 43.40% on the de-CLEF-HIPE and 39.22% on the fr-CLEF-HIPE. We observe a similar effect for the NewsEye datasets. This indicates that organisations are not easy to be distinguished especially for an historical context.

Gold / Predicted	PERS	LOC	ORG	PROD	O
PERS	728	12	6	0	168
LOC	27	359	34	0	33
ORG	3	13	157	9	69
PROD	5	2	6	13	9
O	145	46	41	13	1199

Table 8: Confusion matrix on the fr-CLEF-HIPE test set

2. Named Entity Linking

Digital libraries are composed of a large number of digital contents (e.g., journals, books, magazines, videos, and so on) in several languages about diverse subjects (e.g., history, languages, politics, sciences, philosophy, and so on). Collecting data from different sources leads to revealing the problem of duplicate and ambiguous information about named entities. Therefore, they are often not distinctive since one single name may correspond to several entities. A disambiguation process is thus essential to distinguish named entities to be indexed in digital libraries.

Named Entity Linking (NEL) is the task of recognising and disambiguating named entities to a Knowledge Base (KB). NEL is a challenging task because named entities may have multiple surface forms, such as its full name, partial names, aliases, abbreviations, and alternate spellings [54].

Given a knowledge base containing a set of named entities and a set of documents, the goal of named entity linking is to map each named entity in these documents to its corresponding named entity in a knowledge base (KB) [54], e.g., Wikidata[11]. Wikidata is a free and open knowledge base that can be read and edited by both humans and machines. This KB acts as central storage for the structured data of its Wikimedia sister projects including Wikipedia, Wikivoyage, Wiktionary, Wikisource, and others.

[11] https://www.wikidata.org

In a nutshell, NEL aims to recover the ground truth entities in a KB referred to in a document by locating mentions, and for each mention accurately disambiguating the referent entity (Figure 4).

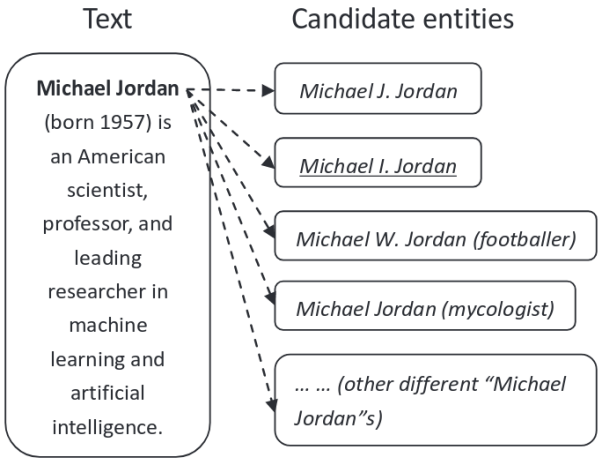


Figure 4: An illustration for the named entity linking task. The named entity mention detected from the text is in bold face; the correct mapping entity is underlined (Shen *et al.* [54]).

Digital libraries often contain the digitised version of old documents that are degraded due to storage conditions, handling of users and inherent vice of the material (e.g., paper naturally deteriorates over time). These problems cause numerous errors at the character and word levels in the OCR of these documents [55]. Consequently, these errors also impact NEL systems by reducing their performance [55].

This section of the deliverable is organised as follows: Section 2.1 makes a brief overview of the most recent and available NEL approaches in the state of the art. Then, we selected the Ganea and Hofman's approach as a baseline [56] (more details in Section 2.2) to be compared with our approach (more details in Section 2.3). We analyse documents in Finnish, French, German, and Swedish languages, and the NEL task provides a link to the Wikidata KB that provides links for an entity in all languages all of Wikipedia that is available. Finally, the experimental setup and the evaluation of this approach are presented in Sections 2.5 and 2.6, respectively.

2.1. An overview of named entity linking

Given a knowledge base containing a set of named entities and a set of documents, the goal of named entity linking is to map each named entity in these documents to its corresponding named entity in a knowledge base [54]. NEL approaches can be divided into two classes:

- Disambiguation approaches: this kind of approach only analyses gold standard named entities in a document and disambiguates them to the correct entry in a given KB.
- End-to-end approaches: this kind of approach processes a document to extract the entities and then disambiguate these extracted entities to the correct entry in a given KB.

Most works in the state of the art are based on three modules: candidate entity generation, candidate entity ranking, and unlinkable mention prediction [54]. More precisely, the first module aims to retrieve related entity mentions in KB that refer to mention in a document. Several works use name dictionary-based techniques [57], surface form expansion from the local document [58], and methods based on search engine [59].

After selecting candidate entities, the second module attempts to rank the most likely link in KB for a mention. Systems in state of the art use mainly supervised and unsupervised methods. These methods consider various techniques to analyse and rank entities, e.g. name string comparison [60], entity popularity [57], entity type [61], textual context [62], and coherence between mapping entities [63]. Finally, the last module validates whether the top-ranked entity identified in the candidate entity ranking module is the target entity for a mention.

Recent neural network methods [56, 64] have established state-of-the-art results, out-performing engineered features based models. These methods combine context-aware word, span and entity embeddings with neural similarity functions.

Next subsections describe the relevant and available NEL systems. Subsection 2.1.1 provides a brief description of disambiguation approaches and Subsection 2.1.2 focuses on the end-to-end approaches.

2.1.1. Disambiguation approaches

Ganea and Hofmann [56] proposed a deep learning model for joint document-level entity disambiguation[12]. In a nutshell, they embed entities and words in a common vector space and use a neural attention mechanism over local context windows to select words that are informative for the disambiguation decision. Their model contains a conditional random field that collectively disambiguates the mentions in a document (more details in Section 2.2).

Le and Titov [64] treated relations between mentions as latent variables in their neural NEL model[13]. As with other recent approaches to NEL [56], they rely on representation learning and learn embeddings of mentions, contexts, and relations in order to reduce the amount of human expertise required to construct the system and make the analysis more portable across languages and domains.

Raiman and Raiman [65] proposed a system for integrating symbolic knowledge into the reasoning process of a neural network through a type system[14]. They constrain the behaviour to respect the desired symbolic structure, and automatically design the type system without human effort. Their model first uses heuristic search or stochastic optimisation over discrete variables that define a type system informed by an oracle and a learnability heuristic. Then, classifier parameters are fitted using gradient descent.

2.1.2. End-to-end approach

Following the idea of jointly analysing the NER and NEL tasks, Kolitsas *et al.* [66] proposed a neural end-to-end NEL system that jointly discovers and links entities in a text document[15]. Their model replaces engineered features by neural embeddings. They first generate all possible spans (mentions) that have at least one possible entity candidate. Then, each mention-candidate pair receives a context-aware

[12]The code is publicly available: https://github.com/dalab/deep-ed
[13]The code is publicly available: https://github.com/lephong/mulrel-nel
[14]The code is publicly available: https://github.com/openai/deeptype
[15]The code is publicly available: https://github.com/dalab/end2end_neural_el

compatibility score based on word and entity embeddings coupled with neural attention and a global voting mechanism (more details are provided in Section 2.3).

Extending this monolingual analysis, cross-lingual named entity linking (XEL) analyses documents and named entities that are in a different language than that used for the content of the knowledge base. In this context, McNamee et al. [67] proposed an XEL approach and examined the importance of transliteration, the utility of cross-language information retrieval, and the potential benefit of multilingual named entity recognition on the XEL task.

Zhou, Rijhwani, and Neubig [68] extensively evaluated the effect of resource restrictions on existing XEL methods in low-resource settings. They investigated a hybrid candidate generation method, combining existing lookup-based and neural candidate generation methods and proposed a set of entity disambiguation features that are entirely language-agnostic. Finally, they designed a non-linear feature combination method, which makes it possible to combine features in a more flexible way.

2.2. Baseline

Ganea and Hofmann [56] proposed a deep learning model for joint document-level entity disambiguation[16] (depicted in Figure 5). They project entities and words in a common vector space, which avoids hand-engineered features, multiple disambiguation steps, or the need for additional ad-hoc heuristics when solving the ED task. Entities for each mention are locally scored based on cosine similarity with the respective document embedding. Combined with these embeddings, they proposed an attention mechanism over local context windows to select words that are informative for the disambiguation decision. The final local scores are based on the combination of the resulting context-based entity scores and a mention-entity prior. Finally, mentions in a document are resolved jointly by using a conditional random field in conjunction with an inference scheme.

Most datasets for NEL are available only in English. Among them, the AIDA data [69] set is the main data used to train NEL systems on the state of the art. Unfortunately, there are few or no datasets for NewsEye languages.

In order to use the Ganea and Hofmann's (GH) system [56] to link mentions from documents in Finnish, French, German, and Swedish, we made some modifications to their approach for linking mentions from OCRed documents [55]. Instead of using the word2vec embeddings, we used the pre-trained multilingual MUSE embeddings[17] [70]. These embeddings are available in 30 languages (including Finnish, French, and German) and they are aligned in a single vector space. Therefore, words like "house" and "talo" ("house" in Finnish) have similar word representations. One of the main goals of using these embeddings is to generate multilingual entity embeddings that can provide entity representations for mentions in several languages. Then, the Ganea and Hofmann's approach will be able to analyse documents in the languages of these embeddings and link them to an English KB. Therefore, we generate the entity embeddings using the English version of Wikipedia and train this system on the AIDA dataset using the MUSE embeddings. In this scenario, the GH's approach can analyse documents in several languages and links their mentions to the English Wikipedia KB.

After obtaining the ID of English pages, we provide the corresponding Wikidata ID for these English pages of Wikipedia.

[16] The code is publicly available: https://github.com/dalab/deep-ed
[17] The MUSE embeddings are available at: https://github.com/facebookresearch/MUSE

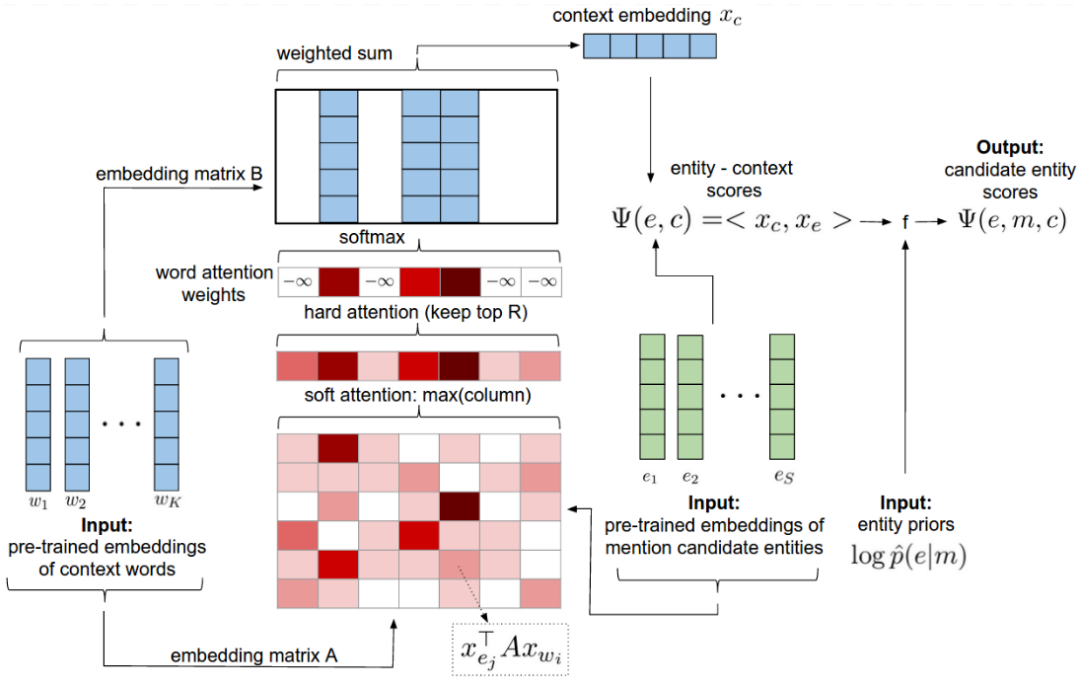


Figure 5: Architecture of the Ganea and Hofmann's approach. Their method uses a local model with neural attention to process context word vectors, candidate entity priors, and embeddings to generate the candidate entity scores [56].

2.3. Multilingual end-to-end entity linking

The NewsEye project aims to analyse historical documents in Finnish, French, German, and Swedish. After recognising NEs (Section 1) in these documents, we disambiguate these entities to a KB. To have a large number of entities for each language, we built a KB and a dataset to train entity embeddings for each language of the project (Section 2.3.1) and, then, we used our entity disambiguation approach (Section 2.3.3) to link these entities for their corresponding language version KB[18].

2.3.1. Building resources

Wikipedia is a multilingual knowledge base (285 languages) with rich information about entities in several languages. From this knowledge base, we can extract several relevant information about entities for the NEL task (contexts, surface names and entity disambiguation cases). Most works in the state of the art use the English version of the Wikipedia as a KB to disambiguate mentions [56, 66]. However, the English Wikipedia may contain fewer pages about persons, organisations, and locations about France and its culture than the French version of Wikipedia. And the same holds for every country and its languages.

Wikipedia has been used to disambiguate mentions in contemporary and historical news documents [56, 66, 71]. Agirre et al.[71] investigated the feasibility of finding matching articles in Wikipedia for a given cultural heritage item in the Europeana corpora. Their results indicated that a substantial number of

[18]The source code will be available at https://github.com/NewsEye/Named-Entity-Linking/tree/master/multilingual_entity_linking.

items (22% of items in Europeana) can be effectively linked to their corresponding Wikipedia article[19]. Other works [72, 73, 74] in the state of the art used the DBpedia KB, which contains structured content extracted from the Wikipedia project. Besides, recent historical datasets (CLEF-HIPE[20] and NewsEye datasets) were annotated with referent URIs taken from Wikidata, which contains structured data of Wikipedia.

In the context of multilingual historical newspapers, documents tend to contain local information that is often specific to a language and one or more related geographical areas. Therefore, the use of knowledge bases in the language of the historical newspaper is an obvious choice, and we disambiguate the entities of historical newspapers to the Wikipedia KB in the corresponding language.

We build a KB for each NewsEye language to have a richer KB for each language. Each language's version of KB is created by the following steps:

- Retrieve the last language version of Wikipedia dump.
- Extract titles and ids of Wikipedia pages.
- Extract list of disambiguation pages and redirection pages.
- Calculate the probability that an entity is related to a mention based on the number of times that mention refers to that entity.

We also build a dataset to train entity embeddings for each language. In this case, we use the methodology used by Ganea and Hofmann [56] to create and train entities embeddings based on the Wikipedia dataset.

2.3.2. Entity embeddings

Following the same idea described in [56], we collected word-entity (word w and entity e) co-occurrence counts (w, e) from two sources: (i) the canonical KB description page of the entity (e.g. entity's Wikipedia page in our case), and (ii) the windows of fixed size surrounding mentions of the entity in an annotated corpus. These counts define a practical approximation of the above word-entity conditional distribution. These words are considered to be the "positive" distribution of entity-related words. Then, a sample of words is selected randomly to create a "negative" distribution of words that are unrelated to the entity e. The objective is to move positive word vectors closer to the embeddings of the entity e and move the vectors of random words further away from the embeddings of the entity e (more details in [56]).

2.3.3. Entity disambiguation

For the entity disambiguation, our model is based on Kolitsas et al.'s work [66] that is a neural end-to-end entity linking model (Figure 6). This model is interesting because we can analyse the entity linking and disambiguation with the same model. Besides, this end-to-end model does not require engineered features, making it easy to upgrade and extend to other languages.

[19] Europeana is composed of a vast number of items; therefore, 22% of items represents a remarkable number of available links to Wikipedia KB.
[20] https://impresso.github.io/CLEF-HIPE-2020/datasets.html

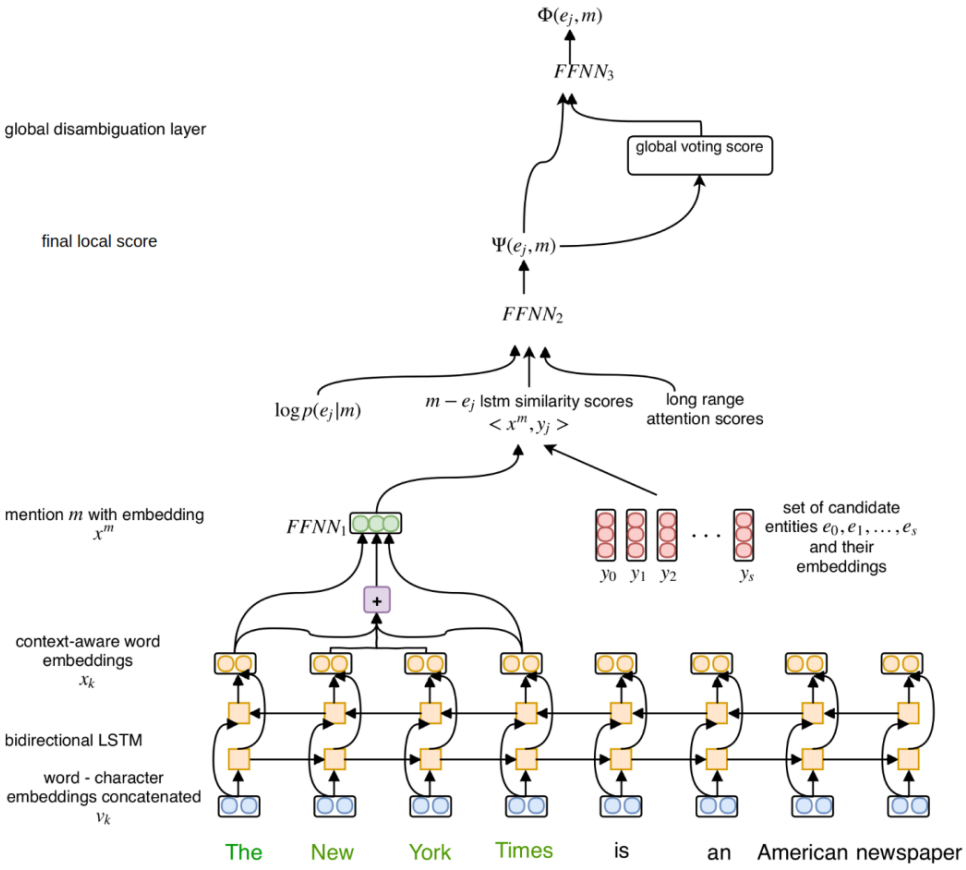


Figure 6: Global model architecture shown for the mention The New York Times. The final score is used for both the mention linking and entity disambiguation decisions (Kolitsas *et al.* [66]).

The first step in the entity linking is to recognise all mentions in a document. Kolitsas et al. used an empirical probabilistic entity-map $p(e|m)$[21] to analyse each span m and select top entities that might be referred by this mention in $p(e|m)$.

Word and character embeddings are concatenated and fed into a BiLSTM to represent a document. This representation is used to project mentions of this document in a dimensional space with the same size of entity embeddings. Entity embedding is calculated separately for each entity using the following exponential model that approximates the empirical conditional word-entity distribution $\widehat{p}(w|e)$ obtained from co-occurrence counts (Section 2.3.2).

In order to analyse long context dependencies of mentions, they used the attention model of GH that gives one context embedding per mention based on informative context words that are related to at least one of the candidate entities. Next, the final local score for each mention is determined by the combination of the $\log p(e|m)$, the similarity between the analysed mention and each candidate entity embeddings, and the long-range context attention for this mention. Finally, a top layer in the neural network promotes the coherence among disambiguated entities inside the same document.

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

2.4. Resources

To the best of our knowledge, there are few publicly available corpora in the literature that are addressed to historical documents. Most NEL corpora are composed of contemporary documents that do not contain the same linguistic variation and OCR problems presented in historical documents.

In order to analyse the robustness of our approach and the state of the art on NEL, we trained the NEL approaches on several types of datasets: news documents (AIDA), Wikipedia documents (WikiANN), and historical documents (CLEF-HIPE and NewsEye NE dataset v1). Then, we analysed the performance of these NEL approaches on historical datasets.

2.4.1. AIDA

The AIDA-CoNLL dataset [69] is based on CoNLL 2003 data that was used for NER task. This dataset is divided into AIDA-train for training, AIDA-A for validation, and AIDA-B for testing. This dataset contains 1,393 Reuters news articles and 27,817 linkable mentions.

2.4.2. WikiANN

Wikipedia is a multilingual resource that currently hosts 294 languages and contains annotated markups and rich informational structures through crowd-sourcing. In this resource, name mentions are often labelled as anchor links to their corresponding referent pages Pan et al. [75]. Taking advantage of this feature, Pan et al. [75] developed an independent language framework to automatically extract name mentions from Wikipedia articles in 282 languages and link them to the English Wikipedia (WikiANN dataset). It is important to note that this dataset is automatically built and that it does contain all the types of named entities used in NewsEye. However, it is an extremely useful resource since it contains datasets in numerous languages, notably in all of the languages of the NewsEye project.

We used the WikiANN on Finnish, French, German, and Swedish. We also converted the links of the English Wikipedia of these datasets for the corresponding language version of the Wikipedia KB, e.g. the French dataset contains links to the French version of Wikipedia KB (Table 9). WikiANN datasets have different numbers of available entities for each language version of the Wikipedia KB. Indeed, some entities presented in the English version of the Wikipedia KB do not have a corresponding entity in the other language versions. When an entity does not exist in a KB, we replace its link with a NIL entry [76]. We do not keep the English identifiers for other languages because they are not consistent between the different language versions of Wikipedia. For example, Wikipedia ID 17515 has different pages for English (“Luxembourg") and Finnish (“Kyberavaruus").

2.4.3. CLEF-HIPE data

Annotated historical data for NEL are too scarce. Fortunately, CLEF-HIPE[22] released training and development datasets for historical documents in English, French, and German (see Section 1.3). In

[22] https://impresso.github.io/CLEF-HIPE-2020/datasets.html

dataset	train	dev	test
de-WikiANN	1,262,142	258,081	264,737
fi-WikiANN	237,779	51,864	50,033
fr-WikiANN	975,416	200,135	250,830
sv-WikiANN	1,248,630	277,397	279,914
de-CLEF-HIPE	3,505	–	1,389
fr-CLEF-HIPE	6,456	–	1,339
de-NewsEye	5,992	1,150	1,618
fi-NewsEye	1,758	288	623
fr-NewsEye	6,467	1,596	1,486
sv-NewsEye	1,780	364	300

Table 9: Number of entities on train/dev/test partitions of datasets.

the NewsEye project, we are interested in the French and German documents of this dataset.

We converted the Wikidata links of CLEF-HIPE datasets for the corresponding language version of the Wikipedia KB and for the English version of Wikipedia KB. Table 9 lists the number of links on the CLEF-HIPE dataset. These datasets have different numbers of available entities for each language version of the Wikipedia KB because English, French and German versions of the Wikipedia KB do not have all entities presented in the Wikidata. When a link is not available in a KB, we replace it by a NIL entry. As with WikiANN, we replace non-existent entities in a KB with NIL entries. Table 10 lists the amount of entities with available links in their corresponding language version of Wikipedia for the CLEF-HIPE data.

Datasets	Wikidata	Wikipedia
de-CLEF-HIPE	3,949	3,280
fr-CLEF-HIPE	5,626	4,439
de-NewsEye	5,333	4,492
fi-NewsEye	1,562	1,204
fr-NewsEye	5,807	4,326
sv-NewsEye	1,529	1,257

Table 10: Number of available links on CLEF-HIPE and NeweEye datasets.

2.4.4. NewsEye data

Recently, the NewsEye project produced through WP1 a dataset composed of historical documents in Finnish, French, German, and Swedish (see Section 1.3.1) with NER and NEL annotations. Similar to the CLEF data, we converted the Wikidata links of NewsEye datasets into the the Wikipedia KBs in the corresponding language and in English (see Table 9). We also replaced non-existent entities in a KB with NIL entries (see Table 10).

2.5. Experimental setup

Entity linking aims to connect named entities to external knowledge bases. In order to accomplish this task, we first need to recognise these entities in the documents and, then, disambiguate them to a KB. In this deliverable, we analyse the disambiguation approaches that only analyse gold standard named entities in a document and disambiguate them to the correct entries in a given KB, i.e. NEL systems know the offset of all mentions in the documents.

2.5.1. Training settings

Both Ganea and Hofmann's approach [56] and our contribution are composed of four models (a model by language). For the GH's approach, we followed the same procedure described in our previous work [77]. More precisely, we used the pre-trained multilingual MUSE word vectors with 300 dimensions[23] to train entity embeddings on the Wikipedia (Feb 2014) corpus. Then, we trained their entity disambiguation approach on AIDA training dataset. Finally, we used the transfer learning procedure to tune this model on the WikiANN datasets for the NewsEye languages. More precisely, we optimised the model learned on the AIDA dataset by training this model on the WikiANN datasets for Finnish, French, German and Swedish.

For our multilingual NEL approach, we used the pre-trained FastText words embeddings [78] with 300 dimensions[24] to train entity embeddings for Finnish, French, German and Swedish on the Wikipedia (Jan 2020) corpus. Then, we trained the Kolitsas et al.'s approach [66] on WikiANN training datasets for each language. Next, we tune our French and German models to the CLEF-HIPE dataset by continuing the training of our models on the training CLEF-HIPE datasets. Finally, we tune our models to the NewsEye datasets by continuing the training of our models on the training NewsEye datasets.

2.5.2. Automatic evaluations

As for named entity recognition, the main evaluation measures for entity linking systems are precision, recall, and F1-score (see Section 1.4). Precision is the fraction of correctly linked entity mentions that are generated by a system. Recall takes into account all entity mentions that should be linked and determines how correct linked entity mentions are with regard to total entity mentions that should be linked. Finally, F1-score is defined as the harmonic mean of precision and recall. These measures are calculated on a full corpus (micro-averaging).

For mentions without corresponding entries in the KB, NEL systems have to provide a NIL entry to indicate that these mentions do not have a ground-truth entity in the KB. In addition, we group these mentions without ground-truth that makes reference to the same entity.

2.6. Experimental Assessment

In the context of historical documents, we analyse the performance of NEL systems on the CLEF-HIPE datasets (Table 11). Differently from contemporary datasets, the CLEF-HIPE datasets are composed

[23] https://github.com/facebookresearch/MUSE
[24] https://fasttext.cc/docs/en/crawl-vectors.html

of historical documents (language variations) and contain OCR errors that may change the spelling of mentions and their contexts. These problems degraded the performance of NEL systems.

The tuning procedure using the WikiANN dataset provided a small improvement for German and did not improve the performance of GH's approach for French. The tuning procedure using the WikiANN corpora did not provide a significant improvement over the GH's model trained on the AIDA dataset. Indeed, AIDA dataset is composed of news documents with longer contexts, while the WikiANN datasets are composed of short sentences that are linked to the KB. In this case, mentions in the WikiANN datasets do not contain long contexts and are limited to the analyse of the surface representation of entities and short contexts.

System	French			German		
	P	R	F1	P	R	F1
Ganea and Hofmann [56] (MUSE emb.: AIDA)	63.5	32.0	42.6	56.7	28.3	37.8
Ganea and Hofmann [56] (MUSE emb.: AIDA + WikiANN)	62.9	28.3	42.2	57.7	28.8	38.4
Our contribution	73.6	39.6	**51.5**	55.2	31.2	**39.9**

Table 11: Precision, Recall, and F1-scores for NEL task on the CLEF-HIPE datasets.

Our contribution achieved the best results for the French and German CLEF-HIPE datasets (improvement of 20.9% and 3.9%, respectively). The stronger performance improvement in French is presumably due to the fact that the training data in French contains 84% more linked entities than the training data in German, as can be observed in Table 9. The main reasons for these improvements are the new probability tables for each language and the tuning training on the historical dataset (CLEF-HIPE). The probability tables $p(e|m)$ of the French and the German versions of the Wikipedia provided more information about persons, organisations, and locations for these languages. Indeed, these tables contain a larger number of entities and their surface names than the table generated by the baseline. These tables helped the disambiguation method to find the entities that are more related to a mention. Moreover, the tuning training on the noised data (CLEF-HIPE) helped our system to reduce the impact of OCR problems and language variations on the disambiguation of mentions in historical documents. Indeed, the analysis of word and character embeddings in our system can provide a better analysis of words and overcome small errors generated by OCR engines.

The performance of NEL systems on NewsEye datasets is described in Table 12. Our contribution achieved the best results for Finnish, French, German, and Swedish on the NewsEye datasets. More precisely, our model achieved relative improvements of 867.7%, 79.2%, 101.7%, and 12.6% for Finnish, French, German, and Swedish datasets, respectively. Similar to the CLEF-HIPE dataset, the generation of the probability table $p(e|m)$ for each language, the analysis of character and word embeddings and the tuned training on the noisy data helped improve the performance over the NewsEye dataset.

The main reason for the poor results of the baseline for the Finnish dataset is the poor quality of the probability table $p(e|m)$ generated from the English Wikipedia and the small amount of available Finnish entities in the English Wikipedia. Indeed, for the baseline, the probability table of the provided candidate entities in the English KB covered only 13.3% of mentions in the Finnish dataset. In contrast, this probability table covered 45.9%, 33%, and 55% of mentions in the French, German, and Swedish datasets, respectively. Most of the mentions in the Finnish dataset are not presented in the English

System	Finnish			French			German			Swedish		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Ganea and Hofmann [56] (MUSE emb.: AIDA)	13.2	1.8	3.1	43.0	19.7	27.0	46.2	15.3	23.0	44.8	24.7	31.8
Ganea and Hofmann [56] (MUSE emb.: AIDA + WikiANN)	13.2	1.8	3.1	42.7	19.6	26.8	46.1	15.2	22.9	44.8	24.7	31.8
Our contribution	51.9	21.1	**30.0**	73.4	36.1	**48.4**	74.6	33.7	**46.4**	61.3	25.3	**35.8**

Table 12: Precision, Recall, and F1-scores for NEL task on the NewsEye datasets.

Wikipedia. Therefore, the probability table generated from the Finnish Wikipedia contains no entry for these mentions, which reduced the performance of the NEL baseline approach, notably in terms of recall.

All in all, our approach outperformed the baselines of both the CLEF-HIPE and the NewsEye datasets. As expected, using specific language versions of Wikipedia provided more relevant information for historical entities (such as surface names and context information). Our probability tables $p(e|m)$ and our tuning procedure using historical documents improved the overall performance of our approach, thanks to an analysis of the context of words that better took OCR errors and language variations into account.

3. Conclusions

The present deliverable describes the final version of the named entity recognition and linking systems. Our improvements on NER and NEL increased the performance of our systems on the analysis of documents in NewsEye languages.

For named entity recognition, we showed that effective NER systems on historical data require large manually annotated corpora. However, such corpora are not always available and small datasets do not allow NER systems to achieve satisfactory performances.We concluded also that contemporary data are not adapted to train robust models to process historical data. However, they can be used with the transfer learning technique to improve the results on historical texts. On the NewsEye NE dataset, we obtained an improvement of about 31% on German, 9% on French, 21% on Finnish and 48% on Swedish. On the CLEF-HIPE datasets, we obtained significant relative improvements over the baseline of about 50% and 9% on German and French corpora respectively. Regarding Finnish data (fi-NLF), transfer learning did not allow us to improve NER performance. This is probably due to the fact that the few existing datasets differed too much. However, NewsEye work in year 1 already allowed to improve the state of the art over the fi-NLF collection from 76% to 87.4%, resulting in a relative improvement of 15%. All the transfer learning NER models that allowed us to achieve these results are available in the NewsEye Github repository[25].

For named entity linking, our contributions provided a significant improvement over the baseline on the French and German CLEF-HIPE datasets (20.9% and 3.9%, respectively) and on the Finnish, French, German, and Swedish NewsEye datasets (867.7%, 79.2%, 101.7%, and 12.6%, respectively). In order to improve our tuning procedure, we would like to extend our training procedure to first train our model on the AIDA dataset and, then, tune our model on the WikiANN, CLEF-HIPE 2020, and NewsEye datasets. This procedure can improve the performance of our contribution; however, AIDA dataset is

[25] https://github.com/NewsEye/Named-Entity-Recognition

English, and WikiANN, CLEF-HIPE, and NewsEye are in Finnish, French, German and Swedish. In this case, multilingual word embeddings can be an alternative to train our model on datasets composed of several languages and, consequently, improve our results. The source code of our NEL approach will be available in the NewsEye GitHub repository[26].

While Task T3.1 is formally ending with this deliverable, the tools presented in this report will continue to be used within the NewsEye workflow, and their output to be integrated into the collections in the NewsEye demonstrator. It is important to underline that the relative quality improvements listed in the present report will actually be higher over the NewsEye collections in practice, due to the improved text input produced in Task T2.2 on automatic text recognition (ATR). Indeed, in terms of character error rate, the reported improvement is of 15–23%, as detailed in Deliverable D2.5. This is expected to have positive impact on our NER and NEL approaches and to trigger an even stronger performance improvement over baselines.

The work led in Task T3.1 produced several public results. In addition to source code, our work on NER and NEL over historical newspapers was recognised by the research community with already 3 top-tier publications: one poster paper on the impact of OCR noise on NER performance [2] at the JCDL 2019 conference (ranked A* by CORE[27]), one short paper on cross-lingual NER [77] at the JCDL 2020 conference, and one long paper on the impact of OCR on NEL [55] published at the ICADL 2019 conference (ranked A by CORE), where it received the award of best paper of the conference. Additional publications describing our most recent results are being prepared. Our publications are constantly updated in the NewsEye Zenodo community[28].

[26] https://github.com/NewsEye/Named-Entity-Linking/

[27] CORE is the usual conference and journal classification in the field of computer science, where conferences and journals are ranked as A* (top 4%), A (next 14%), B (next 26%) and C – see http://www.core.edu.au/conference-portal

[28] https://zenodo.org/communities/newseye/

References

[1] David Nadeau and Satoshi Sekine. “A survey of named entity recognition and classification”. In: *Lingvisticae Investigationes* 30.1 (2007), pp. 3–26.

[2] Ahmed Hamdi, Axel Jean-Caurant, Nicolas Sidere, Mickaël Coustaty, and Antoine Doucet. “An Analysis of the Performance of Named Entity Recognition over OCRed Documents”. In: *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE. 2019, pp. 333–334.

[3] Teemu Ruokolainen and Kimmo Kettunen. “À la recherche du nom perdu–searching for named entities with Stanford NER in a Finnish historical newspaper and journal collection”. In: *13th IAPR International Workshop on Document Analysis Systems*. 2018.

[4] Ralph Grishman and Beth Sundheim. “Message understanding conference-6: A brief history”. In: *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*. Vol. 1. 1996.

[5] Masayuki Asahara and Yuji Matsumoto. “Japanese named entity extraction with redundant morphological analysis”. In: *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*. Association for Computational Linguistics. 2003, pp. 8–15.

[6] Andrew Borthwick, John Sterling, Eugene Agichtein, and Ralph Grishman. “NYU: Description of the MENE named entity system as used in MUC-7”. In: *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998*. 1998.

[7] Satoshi Sekine. “NYU: Description of the Japanese NE system used for MET-2”. In: *Proc. of the Seventh Message Understanding Conference (MUC-7*. Citeseer. 1998.

[8] Daniel M. Bikel, Scott Miller, Richard Schwartz, and Ralph Weischedel. “Nymble: a High-Performance Learning Name-finder”. In: *Fifth Conference on Applied Natural Language Processing*. Washington, DC, USA: Association for Computational Linguistics, Mar. 1997, pp. 194–201. DOI: 10.3115/974557.974586. URL: https://www.aclweb.org/anthology/A97-1029.

[9] Andrew McCallum and Wei Li. “Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons”. In: *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*. Association for Computational Linguistics. 2003, pp. 188–191.

[10] Richard Tzong-Han Tsai, Cheng-Lung Sung, Hong-Jie Dai, Hsieh-Chuan Hung, Ting-Yi Sung, and Wen-Lian Hsu. “NERBio: using selected word conjunctions, term normalization, and global patterns to improve biomedical named entity recognition”. In: *BMC bioinformatics*. Vol. 7. 5. BioMed Central. 2006, S11.

[11] Yassine Benajiba and Paolo Rosso. “Arabic named entity recognition using conditional random fields”. In: *Proc. of Workshop on HLT & NLP within the Arabic World, LREC*. Vol. 8. Citeseer. 2008, pp. 143–153.

[12] Michele Filannino, Gavin Brown, and Goran Nenadic. “ManTIME: Temporal expression identification and normalization in the TempEval-3 challenge”. In: *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Atlanta, Georgia, USA: Association for Computational Linguistics, June 2013, pp. 53–57. URL: https://www.aclweb.org/anthology/S13-2009.

[13] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. “Natural language processing (almost) from scratch”. In: *Journal of Machine Learning Research* 12.Aug (2011), pp. 2493–2537.

[14] Franck Dernoncourt, Ji Young Lee, and Peter Szolovits. “NeuroNER: an easy-to-use program for named-entity recognition based on neural networks”. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 97–102. DOI: 10.18653/v1/D17-2017. URL: https://www.aclweb.org/anthology/D17-2017.

[15] Matthew Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. “Semi-supervised sequence tagging with bidirectional language models”. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 1756–1765. DOI: 10.18653/v1/P17-1161. URL: https://www.aclweb.org/anthology/P17-1161.

[16] Yufan Jiang, Chi Hu, Tong Xiao, Chunliang Zhang, and Jingbo Zhu. “Improved Differentiable Architecture Search for Language Modeling and Named Entity Recognition”. 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)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3585–3590. DOI: 10.18653/v1/D19-1367. URL: https://www.aclweb.org/anthology/D19-1367.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://www.aclweb.org/anthology/N19-1423.

[18] Alan Akbik, Duncan Blythe, and Roland Vollgraf. “Contextual string embeddings for sequence labeling”. In: *Proceedings of the 27th International Conference on Computational Linguistics*. 2018, pp. 1638–1649.

[19] Jason PC Chiu and Eric Nichols. “Named entity recognition with bidirectional LSTM-CNNs”. In: *Transactions of the Association for Computational Linguistics* 4 (2016), pp. 357–370.

[20] Liyuan Liu, Jingbo Shang, Xiang Ren, Frank Fangzheng Xu, Huan Gui, Jian Peng, and Jiawei Han. “Empower sequence labeling with task-aware neural language model”. In: *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.

[21] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. “Deep Contextualized Word Representations”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 2227–2237. DOI: 10.18653/v1/N18-1202. URL: https://www.aclweb.org/anthology/N18-1202.

[22] Abbas Ghaddar and Phillippe Langlais. “Robust Lexical Features for Improved Neural Network Named-Entity Recognition”. In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 1896–1907. URL: https://www.aclweb.org/anthology/C18-1161.

[23] Jana Straková, Milan Straka, and Jan Hajic. “Neural Architectures for Nested NER through Linearization”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 5326–5331. DOI: 10.18653/v1/P19-1527. URL: https://www.aclweb.org/anthology/P19-1527.

[24] Xuezhe Ma and Eduard Hovy. “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)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1064–1074. DOI: 10.18653/v1/P16-1101. URL: https://www.aclweb.org/anthology/P16-1101.

[25] Kate Byrne. “Nested named entity recognition in historical archive text”. In: *International Conference on Semantic Computing (ICSC 2007)*. IEEE. 2007, pp. 589–596.

[26] Gregory Crane and Alison Jones. “The challenge of virginia banks: an evaluation of named entity analysis in a 19th-century newspaper collection”. In: *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*. 2006, pp. 31–40.

[27] Kepa Joseba Rodriquez, Mike Bryant, Tobias Blanke, and Magdalena Luszczynska. “Comparison of named entity recognition tools for raw OCR text.” In: *KONVENS*. 2012, pp. 410–414.

[28] Claire Grover, Sharon Givon, Richard Tobin, and Julian Ball. “Named Entity Recognition for Digitised Historical Texts.” In: *LREC*. 2008.

[29] Yoshihiko Gotoh and Steve Renals. “Information extraction from broadcast news”. In: *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences* 358.1769 (2000), pp. 1295–1310.

[30] Diana Maynard, Valentin Tablan, Cristian Ursu, Hamish Cunningham, and Yorick Wilks. “Named entity recognition from diverse text types”. In: *Recent Advances in Natural Language Processing 2001 Conference*. 2001, pp. 257–274.

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

[32] Benoît Favre, Frédéric Béchet, and Pascal Nocéra. “Robust named entity extraction from large spoken archives”. In: *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. 2005, pp. 491–498.

[33] Mohamed Hatmi. “Reconnaissance des entités nommées dans des documents multimodaux”. PhD thesis. UNIVERSITÉ DE NANTES, 2014.

[34] Alan Ritter, Sam Clark, Oren Etzioni, et al. “Named entity recognition in tweets: an experimental study”. In: *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics. 2011, pp. 1524–1534.

[35] David D Palmer and Mari Ostendorf. “Improving information extraction by modeling errors in speech recognizer output”. In: *Proceedings of the first international conference on Human language technology research*. Association for Computational Linguistics. 2001, pp. 1–5.

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

[37] Daniel M Bikel, Richard Schwartz, and Ralph M Weischedel. “An algorithm that learns what’s in a name”. In: *Machine learning* 34.1-3 (1999), pp. 211–231.

[38] Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. “De-identification of patient notes with recurrent neural networks”. In: *Journal of the American Medical Informatics Association* 24.3 (2017), pp. 596–606.

[39] Sepp Hochreiter and Jürgen Schmidhuber. “Long short-term memory”. In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[40] Alex Graves and Jürgen Schmidhuber. “Framewise phoneme classification with bidirectional LSTM and other neural network architectures”. In: *Neural Networks* 18.5-6 (2005), pp. 602–610.

[41] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. “Learning Word Vectors for 157 Languages”. In: *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*. 2018.

[42] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. “Enriching word vectors with subword information”. In: *Transactions of the Association for Computational Linguistics* 5 (2017), pp. 135–146.

[43] Qi Li. “Literature survey: domain adaptation algorithms for natural language processing”. In: *Department of Computer Science The Graduate Center, The City University of New York* (2012), pp. 8–10.

[44] Sinno Jialin Pan and Qiang Yang. “A survey on transfer learning”. In: *IEEE Transactions on knowledge and data engineering* 22.10 (2009), pp. 1345–1359.

[45] Dong Wang and Thomas Fang Zheng. “Transfer learning for speech and language processing”. In: *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. IEEE. 2015, pp. 1225–1237.

[46] John M Giorgi and Gary D Bader. “Transfer learning for biomedical named entity recognition with neural networks”. In: *Bioinformatics* 34.23 (2018), pp. 4087–4094.

[47] Cosmin Stamate, George D Magoulas, and Michael SC Thomas. “Transfer learning approach for financial applications”. In: *UK Workshop on Computational Intelligence (UKCI)*. 2015.

[48] Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R Curran. “Learning multilingual named entity recognition from Wikipedia”. In: *Artificial Intelligence* 194 (2013), pp. 151–175.

[49] Darina Benikova, Chris Biemann, and Marc Reznicek. “NoSta-D Named Entity Annotation for German: Guidelines and Dataset.” In: *LREC*. 2014, pp. 2524–2531.

[50] Teemu Ruokolainen, Pekka Kauppinen, Miikka Silfverberg, and Krister Lindén. “A finnish news corpus for named entity recognition”. In: *Language Resources and Evaluation* (2019), pp. 1–26.

[51] Zhilin Yang, Ruslan Salakhutdinov, and William W. Cohen. “Transfer Learning for Sequence Tagging with Hierarchical Recurrent Networks”. In: (2017). URL: https://openreview.net/pdf?id=ByxpMd9lx.

[52] Jacob Cohen. “A coefficient of agreement for nominal scales”. In: *Educational and psychological measurement* 20.1 (1960), pp. 37–46.

[53] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. “The Stanford CoreNLP Natural Language Processing Toolkit”. In: *Association for Computational Linguistics (ACL) System Demonstrations*. 2014, pp. 55–60. URL: http://www.aclweb.org/anthology/P/P14/P14-5010.

[54] W. Shen, J. Wang, and J. Han. “Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions”. In: *IEEE Transactions on Knowledge and Data Engineering* 27.2 (2015), pp. 443–460. ISSN: 1041-4347. DOI: 10.1109/TKDE.2014.2327028.

[55] Elvys Linhares Pontes, Ahmed Hamdi, Nicolas Sidere, and Antoine Doucet. "Impact of OCR Quality on Named Entity Linking". In: *Digital Libraries at the Crossroads of Digital Information for the Future - 21st International Conference on Asia-Pacific Digital Libraries, ICADL 2019, Kuala Lumpur, Malaysia, November 4-7, 2019, Proceedings*. 2019, pp. 102–115. DOI: 10.1007/978-3-030-34058-2_11. URL: https://doi.org/10.1007/978-3-030-34058-2%5C_11.

[56] Octavian-Eugen Ganea and Thomas Hofmann. "Deep Joint Entity Disambiguation with Local Neural Attention". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, 2017, pp. 2619–2629. DOI: 10.18653/v1/D17-1277. URL: http://aclweb.org/anthology/D17-1277.

[57] Stephen Guo, Ming-Wei Chang, and Emre Kiciman. "To Link or Not to Link? A Study on End-to-End Tweet Entity Linking". In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Atlanta, Georgia: Association for Computational Linguistics, June 2013, pp. 1020–1030. URL: https://www.aclweb.org/anthology/N13-1122.

[58] Wei Zhang, Yan Chuan Sim, Jian Su, and Chew Lim Tan. "Entity Linking with Effective Acronym Expansion, Instance Selection and Topic Modeling". In: *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three*. IJCAI'11. Barcelona, Catalonia, Spain: AAAI Press, 2011, pp. 1909–1914. ISBN: 978-1-57735-515-1. DOI: 10.5591/978-1-57735-516-8/IJCAI11-319. URL: http://dx.doi.org/10.5591/978-1-57735-516-8/IJCAI11-319.

[59] Xianpei Han and Jun Zhao. "NLPR_KBP in TAC 2009 KBP Track: A Two-Stage Method to Entity Linking". In: *In Proceedings of Test Analysis Conference 2009 (TAC 09)*. MIT Press, 1999.

[60] Zhicheng Zheng, Fangtao Li, Minlie Huang, and Xiaoyan Zhu. "Learning to Link Entities with Knowledge Base". In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. HLT '10. Los Angeles, California: Association for Computational Linguistics, 2010, pp. 483–491. ISBN: 1-932432-65-5. URL: http://dl.acm.org/citation.cfm?id=1857999.1858071.

[61] Mark Dredze, Paul McNamee, Delip Rao, Adam Gerber, and Tim Finin. "Entity Disambiguation for Knowledge Base Population". In: *Proceedings of the 23rd International Conference on Computational Linguistics*. COLING '10. Beijing, China: Association for Computational Linguistics, 2010, pp. 277–285. URL: http://dl.acm.org/citation.cfm?id=1873781.1873813.

[62] Yang Li, Chi Wang, Fangqiu Han, Jiawei Han, Dan Roth, and Xifeng Yan. "Mining Evidences for Named Entity Disambiguation". In: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '13. Chicago, Illinois, USA: ACM, 2013, pp. 1070–1078. ISBN: 978-1-4503-2174-7. DOI: 10.1145/2487575.2487681. URL: http://doi.acm.org/10.1145/2487575.2487681.

[63] Silviu Cucerzan. "Large-Scale Named Entity Disambiguation Based on Wikipedia Data". In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 708–716. URL: https://www.aclweb.org/anthology/D07-1074.

[64] Phong Le and Ivan Titov. "Improving Entity Linking by Modeling Latent Relations between Mentions". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, 2018, pp. 1595–1604. URL: http://aclweb.org/anthology/P18-1148.

[65] Jonathan Raiman and Olivier Raiman. “DeepType: Multilingual Entity Linking by Neural Type System Evolution”. In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*. 2018, pp. 5406–5413. URL: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17148.

[66] Nikolaos Kolitsas, Octavian-Eugen Ganea, and Thomas Hofmann. “End-to-End Neural Entity Linking”. In: *Proceedings of the 22nd Conference on Computational Natural Language Learning*. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 519–529. URL: http://aclweb.org/anthology/K18-1050.

[67] Paul McNamee, James Mayfield, Dawn Lawrie, Douglas Oard, and David Doermann. “Cross-Language Entity Linking”. In: *Proceedings of 5th International Joint Conference on Natural Language Processing*. Chiang Mai, Thailand: Asian Federation of Natural Language Processing, 2011, pp. 255–263.

[68] Shuyan Zhou, Shruti Rijhwani, and Graham Neubig. “Towards Zero-resource Cross-lingual Entity Linking”. In: *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 243–252.

[69] Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. “Robust Disambiguation of Named Entities in Text”. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP ’11. Edinburgh, United Kingdom: Association for Computational Linguistics, 2011, pp. 782–792. ISBN: 978-1-937284-11-4. URL: http://dl.acm.org/citation.cfm?id=2145432.2145521.

[70] Guillaume Lample, Alexis Conneau, Marc’Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. “Word translation without parallel data”. In: *International Conference on Learning Representations*. 2018. URL: https://openreview.net/forum?id=H196sainb.

[71] Eneko Agirre, Ander Barrena, Oier Lopez de Lacalle, Aitor Soroa, Samuel Fernando, and Mark Stevenson. “Matching Cultural Heritage items to Wikipedia”. In: *Eight International Conference on Language Resources and Evaluation (LREC)*. 2012. ISBN: 978-2-9517408-7-7.

[72] Francesca Frontini, Carmen Brando, and Jean-Gabriel Ganascia. “Semantic Web Based Named Entity Linking for Digital Humanities and Heritage Texts”. In: *Proceedings of the First International Workshop Semantic Web for Scientific Heritage at the 12th ESWC 2015 Conference*. Vol. 1364. June 2015.

[73] Gary Munnelly and Seamus Lawless. “Investigating Entity Linking in Early English Legal Documents”. In: *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*. JCDL’18. Fort Worth, Texas, USA: Association for Computing Machinery, 2018, pp. 59–68. ISBN: 9781450351782. DOI: 10.1145/3197026.3197055. URL: https://doi.org/10.1145/3197026.3197055.

[74] Seth van Hooland, Max De Wilde, Ruben Verborgh, Thomas Steiner, and Rik Van de Walle. “Exploring entity recognition and disambiguation for cultural heritage collections”. eng. In: *DIGITAL SCHOLARSHIP IN THE HUMANITIES* 30.2 (2015), pp. 262–279. ISSN: 2055-7671. URL: http://dx.doi.org/10.1093/llc/fqt067.

[75] Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. “Cross-lingual Name Tagging and Linking for 282 Languages”. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, 2017, pp. 1946–1958.

[76] Avirup Sil, Gourab Kundu, Radu Florian, and Wael Hamza. “Neural Cross-Lingual Entity Linking”. In: *AAAI*. AAAI Press, 2018, pp. 5464–5472.

[77] Elvys Linhares Pontes, Antoine Doucet, and Jose G. Moreno. “Linking Named Entities across Languages using Multilingual Word Embeddings”. In: *Jointed Conference on Digital Libraries (JCDL) 2020*. 2020.

[78] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. “Enriching Word Vectors with Subword Information”. In: *Transactions of the Association for Computational Linguistics* 5 (2017), pp. 135–146. DOI: 10.1162/tacl_a_00051. URL: https://www.aclweb.org/anthology/Q17-1010.

A. Appendix: Named entity and stance annotation guidelines

NewsEye

Named Entity and Stance Annotation Guidelines

Version: 3.1 - March 2020

Initially based on version 2.0 of the Impresso NE annotation guidelines[1]

[1] By Maud Ehrmann, Camille Watter, Matteo Romanello, Simon Clematide (Camille Watter for initial Quaero translation and impresso adjustments, Maud Ehrmann for reshaping, reformulation and impresso adjustments, Simon Clematide and Matteo Romanello for impresso adjustments).

CC BY 0

1. Preamble

Guidelines genealogy

While the part of the guidelines on stance detection annotation is new, the NewsEye NE annotation guidelines are derived from Impresso NE annotation guidelines which are derived from Quaero guidelines[2]. Originally designed for the annotation of "extended" named entities (i.e. more than the 3 or 4 traditional classes) in French speech transcriptions, Quaero guidelines have furthermore been used on historic press corpora[3]. *Impresso* guidelines main's difference with respect to Quaero's is *reduction*: only a subset of Quaero entity types and components are considered, as well as a subset of linguistic units eligible as named entities. These adaptations result from what we deemed most relevant to annotate in our context, and from time and resource constraints. Despite these adaptations, *impresso* annotated corpora will mostly remain compatible with Quaero guidelines.
Followingly, the NewsEye guidelines are intended to be compatible with the Impresso ones, in order to allow the produced datasets to be compatible too, and so that both projects (and the community at large) can benefit of combined efforts and a significant amount of compatible training data, rather than from independent and incompatible smaller collections.

Application context

The objective is to extract information from historical newspaper articles, in view of supporting the search, filtering and analysis of large collections of newspaper archives, and of building a historical knowledge base, eventually connected to others (e.g. Wikidata, HistHub).
As such, our objective is similar to one of classical media monitoring, where we want to extract salient 'journalistic' entities among the typical '5Ws' (Who, What, Where, When, Why).
Our context is however different in that documents are not contemporary but historical, and final users are not politicians or economic actors but scholars. This led us to some adjustments with respect to, mainly: (a) the tag set (addition of newspaper-related specific types), (b) granularity of annotation (emphasis on *Person* type in view of the biographical scenario), and (c) concrete implementation of annotation (flag for noisy entities, capacity to view the original facsimile).

2. General instructions

2.1 Entity types and subtypes

The objective is to annotate all named mentions in texts, of the following types and subtypes:

[2] See the original Quaero guidelines:
http://www.quaero.org/media/files/bibliographie/quaero-guide-annotation-2011.pdf , and our English translation:
https://docs.google.com/document/d/13LRvP5Qh99myEEH_lqqcHaa3S-nZ2Sr71iZ5YbecDCc/edit#
[3] See ELRA catalog entry: http://catalog.elra.info/en-us/repository/browse/ELRA-W0073/

CC BY

Type	Subtypes
pers	pers.articleauthor
loc	
org	
(prod)	prod.media

In the NewsEye NE annotation, the subtypes are very limited:

- <pers.articleauthor> is a specific subtype of person describing the author of an article, especially useful for newspapers. Every other type of person NE should be simply annotated with <pers>. This is further detailed in Section 3.1
- <prod.media> is a specific subtype of human production described in Section 3.4. As this is the only type of human production we wish to annotate in NewsEye, we will never actually use the <prod> annotation but only the <prod.media> annotation
- <org> and <loc> are for organisations and locations. No subtypes are to be taken into account.

2.2 Named entity mention lexical characteristics

A. Nature.

Linguistic units considered as named entities must include a proper name, or a definite description having the status of a proper name[4]. Although the definition of a proper name is not straightforward, here are a few characteristics commonly accepted (not valid in all cases nor in all languages): presence of majuscule, non inclusion in lexical but in encyclopedic dictionaries, absence of meaning (the name *George* does not carry - per se - any information about the type of entity that can be called this name, while the noun “table” gives specific information about the type of objects that can be called by it - i.e. having a plateau and feets), and absence of compound meaning (the *White House* does not refer to any house which is white, la *Gare de Lyon* is not in Lyon, le *Pont Neuf* is very old).

We do not specify further the definition of proper names[5], but instead rely on the linguistic intuition/awareness of annotators, who should always keep in mind our objective of extracting ‘journalistic’ information typically conveyed via referential entities. There will be borderline cases, which we ask annotators to report in a separate file for further discussion[6].

Phrases such as

- *Die präkolumbianische Zivilisation, la civilisation précolombienne*
- *l’armée bavaroise*

[4] This position is more strict than Quaero, which allow entities to be composed of proper names and of common nouns (cf. Section 1.5 or Quaero guidelines).

[5] A rabbit hole. For an overview of proper name definition see: https://hal.archives-ouvertes.fr/tel-01639190

[6] See the last section “Quick guide and concrete implementation”.

CC BY

- *les forces tchadiennes*
- *le gouvernement français*

are *not* annotated because they do not contain proper names.

Phrases such as:

- *le gouvernement Franco*
  le <org> gouvernement
  <pers> Franco </pers>
  </org>

are annotated.

In front of some definite descriptions, it might be difficult to decide what to do, e.g. *la commission Impériale, l'escadre de Nelson.* In such difficult cases, consider the following:

- definite descriptions which can be considered as named entities tend to have a **nominative function** (like proper names) rather than a descriptive function. What a definite description says literally about a referent is less important than the nominative aspect.
- even though, some named entities are definite descriptions which are descriptive, e.g. "Syndicat National de la Magistrature". In such cases, what makes it a named entity is the **referential stability**: the entity referred to is always the same.
- in general, our bottom line is: **we do not accept borderline definite descriptions.**

B. Boundaries. A named entity can be the head of several nominal syntagms but not all of them are annotated.

- Named entity mentions exclude:
  - subordinate clauses;
  - incidental clauses or insertions : if an insertion divides a mention, each part is annotated separately;
  - determiners.

- Named entity mentions include:
  - pre modifiers
    Le soviétique Alexandre Avreni a déclaré…
    Le compatriote Serge Martin est déçu…
    La grande Armée Rouge

  - post modifiers, including in apposition:
    Anne Hidalgo, maire de Paris, a déclaré
    Anne Hidalgo, une forte femme, a déclaré
    Shekau, chef de l'une des trois factions de Boko Haram et fondateur historique du groupe, diffusait une vidéo…

- Special cases with noisy OCR:

CC BY

When it is difficult to establish the boundary of a mention because of noisy OCR:

- look at the image
- include, in the annotation, the garbage characters which you think should have been recognized and should be part of the mention
- mark the mention with the flag "noisy-entity" and add your OCR hypothesis correction.
  ex: in the string *Trève ** (which stands for *Trèves*), the full string *Trève ** should be annotated, not only *Trève.*

- Special case with German compounds:
  Apply the cross-lingual or decomposition test, i.e. translate the compound to French and in the German compound annotate only what should be annotated in French.

  Baslerpropaganda
  => French translation (decomposition): propagande baloise
  => no annotation

  Zürichputsch
  => French translation (decomposition): le putsh de Zurich (Putsch von Zürich)
  => annotation of "Zürich"
  <loc>Zürich</loc>putsch

  Donaufestungen
  => Festungen an der Donau
  => annotation of "Donau"
  <loc>Donau</loc>festungen

  Der am Montag in Kairo ermordete ägyptische Ministerpräsident Al-Nokraschi
  => "Le premier ministre égyptien Al-Nokraschi, qui a été assassiné au Caire lundi, ..."
  <pers>ägyptische Ministerpräsident Al-Nokraschi</pers>

The connecting "s" in German compounds is *not* annotated:
Völkerbundsmitgliedern
=> only *Völkerbund* is annotated
<org>Völkerbund</org>smitgliedern

2.3 Nesting and special constructions

A. Nested entities. An entity can be nested in another entity or in an entity component.

CC BY 4

- nested entities are annotated for the types PERS, LOC, ORG, with a limit of nested entities of depth 1, i.e. a nested entity cannot contain a nested entity (note that entity linking and stances are not concerned by nested entities).

La Feuille d'Avis de Neuchâtel
<prod.media>Feuille d'Avis de
<loc>Neuchatel</loc>
</prod.media>

La société du Parc du Creux-du-Vent...
<org>société du
<loc>Parc du Creux-du-Vent</loc>
</org>

Le maire de Paris Bertrand Delanoë a déclaré
<pers>
maire de <loc>Paris </loc> Bertrand Delanoë
</pers>

dem Preussischen Staatsminister der auswärtigen AngelegenHeiten, Graf von Goltz
<pers>
Preussischen
Staatsminister der auswärtigen Angelegenheiten
Grafvon Goltz
</pers>

- components of nested entities are not annotated

B. Coordination. Entities coordinated based on a common descriptor or trigger word are annotated separately. Type is inferred from the type of the coordinated entity. Coordinating conjunctions are excluded from annotation.

Der Bodensee, Starnberger See und Müritz
Der <loc> Bodensee </loc>,
<loc> Starnberger See </loc>, und
<loc> Müritz </loc>

vallées de la Lorraine, de l' Alsace et de la Champagne
<loc> vallées de la Lorraine</loc>,
<loc> de l'Alsace </loc> et
<loc> de la Champagne </loc>

In any cases, a proper name must be present in the entity mention, therefore only one entity is annotated when it is not the mentions but the title/trigger words which are coordinated:

CC BY 5

Monsieur et Madame Chirac…
Monsieur et <pers>Madame Chirac </pers> …

Ost und Mitteleuropa….
Ost und <loc>Mitteleuropa</loc>....

Special case of a coordination within a component: this produces 2 separate components, excluding the coordination.

Shekau, chef de l'une des trois factions de Boko Haram et fondateur historique du groupe, diffusait une vidéo…
<pers> Shekau,
chef de l'une des trois factions de
<org>Boko Haram<org> et
fondateur historique du groupe
</pers>

C. Elaboration. When a mention is complemented with an acronym or an abbreviation, both are treated as distinct entities.

DAISY das dynamische Auskunfts- und Informationssystem
<org> DAISY </org> das
<org> Dynamische Auskunfts- und Informationssystem </org>

Agipi association d'assurés pour la prévoyance, la dépendance et l'épargne-retraite
<org> Agipi</org>\
<org> Association d'assurés pour la prévoyance , la dépendance et l'épargne-retraite
</org>

D. Difficult example(s)

der bekannte Irländer Theobald Wolfe Tone, den man auf….
<pers>
bekannte
Irländer
Theobald Wolfe Tone
</pers>

CC BY 6

2.4 Ambiguities

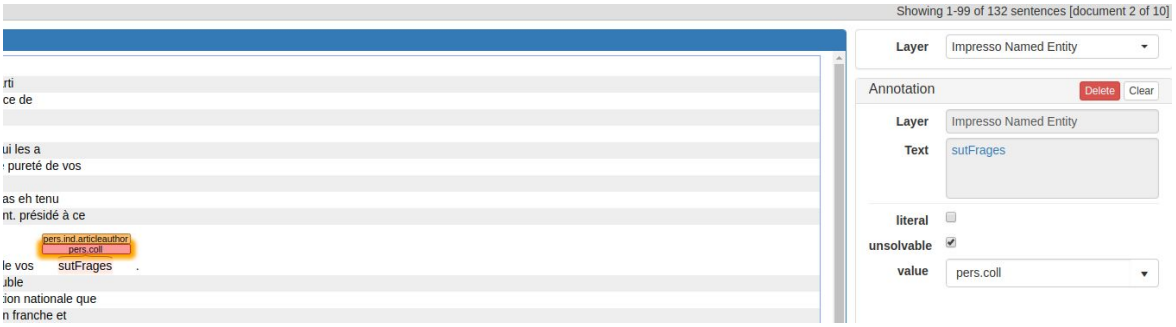
A. Unsolvable ambiguities: flag 'unsolvable'

Even in context, some entities can remain ambiguous:

<??>Yves Rocher</??> lässt sich in Vannes nieder

<??>Yves Rocher</??> va s'installer à Vannes

In these cases, the annotation is 'double' and includes 2 types. To differentiate this annotation from a metonymic one (which also results in two tags for one mention), annotator should add the flag 'unsolvable' to one of the 2 annotations.



In case of unsolvable ambiguity, it is mandatory to indicate 2 types minimum.

B. Metonymy.

Metonymy is a figure of speech in which a thing or a concept is not called by its own name but by the name of something intimately associated to that thing or concept. The category to which the mentioned entity inherently belongs is annotated and is nested within the category that the term refers to in the context.

In Inception annotation tool, the literal annotation has to be flagged with the corresponding flag.

Eine Erklärung des Quai d'Orsay
Eine Erklärung des
<org>
<loc> Quai d'Orsay </loc>
</org>

Une déclaration du Quai d'Orsay
Une déclaration du
<org>
<loc> Quai d'Orsay </loc>
</org>

Die rue de Grenelle hat auf diese Aussage reagiert
Die
<org>
<loc> rue de Grenelle </loc> </org> hat auf diese Aussage reagiert

La rue de Grenelle a réagi à cette déclaration
La
<org>
<loc> rue de Grenelle</loc>
</org> *a réagi* à cette déclaration

Die Élysée erklärt
Die <org><loc> Élysée </loc></org> erklärt

L'Élysée a déclaré...
L' <org><loc>Élysée </loc></org> a déclaré...

CC BY

3. Entities

3.1 Person

When the entity refers to individual or collective person (more than one individual) including fictitious persons. Even in the case of a collective person annotation, there must be the presence of a proper name (e.g. *the Beatles, the Cohen Brothers, die Habsburger, les Bourbons*).

A. Subtype

- `pers.articleauthor`: special type to recognize authors of newspaper articles, either full names or initials at the end of the text, or within a formula such as "*from or correspondant xx in yy*"
  This is the only subtype for persons, every other person NE is annotated with <pers>

  Following expressions are **not** annotated:

  die französischen Opfer des Unfalls, die chinesischen Touristen / les victimes françaises de l'accident, les voyageurs chinois

  Die Maya Zivilisation / la civilisation Maya

  Arbeiter, Menschen, die Verletzten; */ le monde ouvrier, les êtres humains, les blessés, etc.*

  Die Protestanten, die Spanier / les protestants, les espagnols

B. Coverage of the type Person

- Considered as Person:
  - real persons
  - imaginary characters and characters of literature pieces (e. g. *Asterix,* when referring to the character, but not when referring to the work e.g. *Uderzo ist der Schöpfer der Comic-Reihe Asterix, Uderzo est le créateur de la BD Astérix*)
  - religious figures (*God*)

- Not considered as Person:
  - expressions which do not contain a proper name
  - demonyms which do not modify a proper name:
    e.g. Le français s'est classé quatrième.
    Der Schweizer ist Vierter geworden
  - isolated functions not attached to a person name
  - religious persons are not annotated in namedays and addresses

Der Bürgermeister von Paris => only 'Paris'	*le maire de Paris* => only 'Paris'

CC BY

Die Bürgermeister von Frankreich => only 'France' *Der Forscher des CNRS* => only 'CNRS' *Der Präfekt ist essen gegangen* => no annotation *Angelegenheiten* => no annotation *Ein britischer Journalist* => no annotation *Der ehemalige Bürgermeister von Paris* => only 'Paris' *Die Polizisten* => no annotation *Die Polizisten von Paris* => only 'Paris' *Präsident der Republik* => no annotation *Präsident der islamische Republik Pakistan* => only `Pakistan' *Einer der Polizisten* => no annotation *Ex Miss Italien* => no annotation *Der Papst* => no annotation	*les maires de France* => only 'France' *le chercheur CNRS* => only 'CNRS' *le préfet est parti manger* => no annotation *un journaliste britannique*=> no annotation *l'ancien maire de Paris* => only 'Paris' *les pompiers*=> no annotation *les pompiers de Paris* => only 'Paris' *président de la république*=> no annotation *président de la République islamique du Pakistan* => annotate only `Pakistan' *l'un des pompiers*=> no annotation *ex Miss Italie* => no annotation *le Pape* => no annotation *la saint Nicolas* => no annotation

func / title / name	
Seine Königliche Hoheit Prinz Rainier <pers> Seine Königliche Hoheit Prinz Rainier </pers>	*Son Altesse Royale le prince Rainier* <pers> Son Altesse Royale le prince Rainier </pers>
Der König Mohamed VI Der <pers> König Mohamed VI </pers>	*Le roi Mohamed VI* Le <pers> roi Mohamed VI </pers>
Ihr Majestät der König Mohamed VI <pers> Ihre Majestät der König Mohamed VI </pers>	*Sa Majesté le roi Mohamed VI* <pers> Sa Majesté le roi Mohamed VI </pers>
Der Dr. Duboc, ehemaliger Abteilungsleiter von Pitié-Salpêtrière Der <pers> Dr. Duboc ehemaliger Abteilungsleiter von Pitié-Salpêtrière </pers>	*Le Dr. Duboc, ancien chef de service à la Pitié-Salpêtrière* Le <pers> Dr. Duboc ancien chef de service à la Pitié-Salpêtrière </pers>

CC BY

Der Bürgermeister Delanoë Der <pers> Bürgermeister Delanoë </pers>	*Le maire Delanoë* Der <pers> maire Delanoë </pers>
Bertrand Delanoë, der Bürgermeister von Paris <pers> Bertrand Delanoë, Der Bürgermeister von <loc> Paris </loc> </pers>	*Bertrand Delanoë, le maire de Paris* <pers> Bertrand Delanoë, le maire de <loc> Paris </loc> </pers>
Herr Martin, der türkische Botschafter in Frankreich <pers> Herr Martin, der türkische Botschafter in <loc> Frankreich </loc> </pers>	*Monsieur Martin, l'ambassadeur de Turquie en France* <pers> Monsieur Martin </name>, l'ambassadeur de <loc> Turquie </loc> en <loc> France </loc> </pers>
General De Gaulle <pers> General De Gaulle </pers>	*le général De Gaulle* Le <pers> Général De Gaulle </pers>
qualifier	
Der konservative Christoph Blocher Der <pers> konservative Christoph Blocher </pers>	*Le socialiste Bertrand Delanoë* Le <pers> socialiste Bertrand Delanoë </pers>
name	
von Lange <pers> von Lange </pers>	*De Gaulle* <pers> De Gaulle </pers>
demonym	
Der Engländer Tony Blair erklärt…. Der <pers> Engländer Tony Blair </pers>	*L'anglais Tony Blair a déclaré….* L' <pers> anglais Tony Blair </pers>

CC BY

3.2 Organisations

Examples of organisations

- A company which sells products or provides services that are not only administrative. It includes both private and public companies, as well as hospitals, schools, universities, political parties, trade unions, police, gendarmerie, churches, (named) armies, sportive clubs, etc.

Die Peugeot Gesellschaft
```
Die <org>
Peugeot Gesellschaft </org>
```

Ich arbeite bei Peugeot
```
Ich arbeite bei
<org> Peugeot
</org>
```

Die UNESCO
```
Die <org> UNESCO
</org>
```

Die Rote Armee
```
Das <org> Rote Armee </org>
```

Die Grüne Partei: 'Partei' is part of the name of this party (GPS)
```
Die <org> Grüne Partei</org>
```

Die Partei JungsozialistInnen Schweiz: 'Partei' is not part of the name of this party (juso)
```
Die <org> Partei
JungsozialistInnen Schweiz
</org>
```

Die Gewerkschaft UNIA
```
die <org>
     Gewerkschaft
    UNIA
</org>
```

Die Gewerkschaft des Verkehrspersonals
```
Die <org> Gewerkschaft des
Verkehrspersonals </org>
```

La société Peugeot
```
La <org>société Peugeot</org>
```

Je travaille chez Peugeot
```
Je travaille chez
<org> Peugeot</org>
```

L' UNESCO
```
L' <org> UNESCO</org>
```

L'Armée Rouge
```
L' <org> Armée Rouge</org>
```

l'hôpital d'instruction des armées du Val-de-Grâce
```
L' <org> hôpital d'instruction
des armées du Val-de-Grâce
</org>
```

Le parti socialiste: 'parti' is part of the name of this party (PS)
```
Le <org> parti socialiste
</org>
```

Le parti Europe Écologie: 'parti' is not part of the name of this party (EE)
```
Le  <org> parti Europe Écologie
</org>
```

Le syndicat FSU
```
Le <org> syndicat FSU </org>
```

Le syndicat national de la magistrature
```
Le <org> syndicat national de
la magistrature </org>
```

- An organisation which plays a mainly administrative role. It is often an administrative and/or geographical division. This includes town halls, city council, regional council, state council,

CC BY

federal council, named government, ministry parliament, prefectures, ministries dioceses, tribunal, court, government treasury, public treasury, international org.

Die Stadtverwaltung Bern
```
Die <org>
Stadtverwaltung
<loc> Bern </loc>
</org>
```

La Mairie de Paris
```
La <org> mairie de
    <loc> Paris </loc>
</org>
```

Das Bistum Basel
```
Das <org> Bistum
    <loc> Basel </loc>
</org>
```

Le diocèse de Blois
```
Le <org> diocèse de
    <loc> Blois </loc>
</org>
```

3.3 Locations

Examples of locations, all instinctively marked as <loc>

A. Administrative locations: refer to a territory with a geopolitical border.

- district, city: includes cities and all smaller units:
  - city, village, hamlet, locality, commune;
  - part of the city: district, borough, etc.

Zürich
```
<loc> Zürich </loc>
```

Paris
```
<loc> Paris </loc>
```

La Bollline
```
<loc> La Bolline </loc>
```

Val de Crüye
```
<loc> Val de Crüye </loc>
```

Maison Blanche
```
<loc> Maison Blanche </loc>
```

La ville de Paris
```
La <loc> ville de Paris </loc>
```

Der Kreis 4
```
Der <loc>Kreis 4 </loc>
```

Die Stadt Zürich
```
Die <loc> Stadt Zürich</loc>
```

Big Apple
```
<loc> Big Apple</loc>
```

Le 13e arrondissement
```
Le <loc> 13e arrondissement </loc>
```

La ville rose
```
La <loc> ville rose </loc>
```

CC BY

- region: refers to internal divisions within a state and includes all units between country and city levels: administrative and traditional regions, departments, counties, departmental districts, Swiss cantons, including the associated municipalities communities of municipalities, urban communities, etc.

Die Autonome Gemeinschaft Baskenland
```
Die <loc> Autonome Gemeinschaft
Baskenland </loc>
```

Im Süden von Israel
```
<loc>
    Im Süden  von
   Israel
</loc>
```

la CAPS
```
la <loc> CAPS </loc>
```

Au sud d'Israël
```
au <loc>
sud d'Israël
</loc>
```

Le Pays basque espagnol
```
Le <loc> Pays basque espagnol
</loc>
```

- national: for countries.

Die Schweiz, Vereinigtes Königreich, die Vereinigten Staaten, Andorra; Monaco, la France, le Royaume-Uni, les États-Unis.

Das Vereinigte Königreich
```
Das <loc> Vereinigte Königreich
</loc>
```

Le Royaume-Uni
```
Le <loc> Royaume-Uni </loc>
```

- supranational: refers to world regions, continents, etc. :

Der Nahe Osten, das Baskenland, Katalonien, der Commonwealth, der Norden, le Moyen Orient;
le Pays basque, la Catalogne, le Commonwealth, l'Afrique subsaharienne, le Sud [7]

Das Baskenland
```
Das <loc> Baskenland </loc>
```

Die Region um den Atlas
```
Die <loc>
   Region um den Atlas
</loc>
```

Le Pays basque
```
Le <loc> Pays basque </loc>
```

La région de l'Atlas
```
La <loc>
     Région de l'Atlas
</loc>
```

B. Physical places:

[7] In the sense of the countries of the South. In other contexts, the south could designate other geographical locations (*le Sud de la France*).

CC BY

- terrestrial physical locations:
  Geonyms[8] include names given to natural geographical spaces, such as deserts, mountains, mountain chains, glaciers, plains, chasms, plateaus, valleys, volcanoes, canyons, etc.

Der Ätna
```
Der <loc> Ätna </loc>
```

L'Etna
```
L' <loc> Etna </loc>
```

Die Wüste Gobi
```
Die <loc>
Wüste Gobi
</loc>
```

Le desert de Gobi
```
Le <loc>
désert de Gobi
</loc>
```

- aquatic physical sites:
  Hydronyms [9] refer to water bodies [10], such as rivers, streams, ponds, marshes, lakes, seas, oceans, marine currents, canals, springs, etc.

Die Spree
```
Die <loc> Spree </loc>
```

La Seine
```
La <loc> Seine </loc>
```

Der Canal Saint-Martin
```
Der <loc>
  Canal Saint-Martin
</loc>
```

Le Canal Saint-Martin
```
Le <loc>
  Canal Saint-Martin
</loc>
```

- astronomical physical places: includes planets, stars, galaxies, etc., and their parts.

Der Mond
```
Der <loc> Mond </loc>
```

La Lune
```
La <loc> Lune </loc>
```

Die Milchstrasse
```
Die <loc> Milchstrasse </loc>
```

la mer de la tranquillité
```
La <loc> mer de la tranquillité
</loc>
```

C. Pathways:
refer to streets, squares, roads, highways, etc.

Die Autobahn A6
```
Die <loc>
Autobahn A6
</loc>
```

place de l'Abbé Georges Hénocque
```
<loc> place de l'
```

[8] Definition taken from Mickaël Tran's thesis, Université de Tours, 2006, p. 84
[9] Definition taken from Mickaël Tran's thesis, Université de Tours, 2006, p. 84
[10] We include water streams as well.

CC BY

Die A6
```
Die
<loc> A6
</loc>
```

Die Nordring Autobahn
```
Die <loc>
 Nordring Autobahn
</loc>
```

Der Nordring
```
Der <loc>
    Nordring
</loc>
```

```
 <pers> Abbé Georges Hénocque
</pers>
</loc>
```

rue de Vaugirard (Vaugirard is a village)
```
<loc> rue de
Vaugirard
</loc>
```

la 118
```
la <loc> 118 </loc>
```

le triangle de Rocquencourt
```
le <loc> triangle de
Rocquencourt </loc>
```

L'autoroute A6
```
L' <loc> autoroute A6 </loc>
```

rue des Glycines
```
<loc> rue des Glycines </loc>
```

D. Buildings :

Named buildings (train station, museum, ..) as well as their extensions (stadium, campus, university, camping...) often refer to the physical location of an organisation.

Zürich Hauptbahnhof
```
<loc> Zürich Hauptbahnhof </loc>
```

Bern Bümpliz Nord
```
<loc> Bern Bümpliz Nord
</loc>
```

Der ehemalige Bahnhof Letten
```
Der <loc>
ehemalige Bahnhof Letten
</loc>
```

Schloss Kyburg
```
<loc>
  Schloss Kyburg
</loc>
```

Die Kyburg
```
Die <loc> Kyburg </loc>
```

La gare de Rungis
```
La <loc> gare  de Rungis </loc>
```

la gare Saint-Germain Grande Ceinture
```
la <loc> gare Saint-Germain
Grande Ceinture </loc>
```

l'ancienne gare de Rungis
```
l' <loc>
ancienne gare de Rungis
</loc>
```

le palais de l'Élysée
```
Le <loc>
    palais de l'Élysée
</loc>
```

l'Élysée
```
l' <loc>  Élysée </loc>
```

E. Addresses:

- physical addresses: an address is a point in space (e.g. a point in a street)

CC BY

Ich wohne in der Shilstrasse 15 3. Stock

```
Ich wohne in der
<loc>
    Sihlstrasse 15 3. Stock
</loc>
```

9 place de Rungis

```
<loc>
  9 place de Rungis
</loc>
```

J'habite 15 rue de Vaugirard escalier 2

```
J' habite
<loc>
    15 rue de Vaugirard escalier 2
</loc>
```

31, Quai du Mont-Blanc Genova

```
<loc>
  31, Quai du Mont-Blanc Genova
</loc>
```

- electronic addresses:
  Electronic coordinates: a telephone or fax number, url, E-Mail address, frequency radio, social network identifiers (*Facebook, Twitter*) or tools for internet communication (*Skype*), etc.

Meine Nummer lautet 01 69 85 80 02

```
Meine Nummer lautet
<loc> 01 69 85 80 02 </loc>
```

Mein Skype-Name ist jean.dupont

```
Mein Skype-Name ist
<loc> jean.dupont </loc>
```

Radio Bleue auf 98.8 MHz

```
<prod.media> Radio Bleue
</prod.media> auf
<loc> 98.8 MHz </loc>
```

Folgt mir auf Twitter unter @leguidedannotation

```
Folgt mir auf Twitter unter
<loc>
  \@leguidedannotation
</loc>
```

mon numéro est le 01 69 85 80 02

```
mon numéro est le
<loc> 01 69 85 80 02 </loc>
```

mon identifiant skype est jean.dupont

```
mon identifiant skype est
<loc> jean.dupont </loc>
```

Radio Bleue sur 98.8 MHz

```
<prod.media> Radio Bleue
</prod.media> sur
<loc> 98.8 MHz </loc>
```

suivez-moi sur Twitter à @leguidedannotation

```
suivez-moi sur Twitter à
<loc>
  \@leguidedannotation
</loc>
```

CC BY

3.4 Human production

Media (to annotate as <prod.media>): newspapers, magazines, broadcasts, sales catalogues, etc.
(*Die Zeit;* Le *Figaro*, *Le sept à huit, La ferme célébrités*).
The name of the newspaper under annotation is annotated only when it is mentioned in the body of the newspaper articles. It is not annotated at the page-level including advertisements, images, footers, headers...

Doctrine (to ignore): political, philosophical, religious, sectarian doctrines.
(Der *Sozialismus, Theravada Buddhismus; Zeugen Jehovas;* Le *socialism*, le *bouddhisme theravâda*,le *structuralism*, la *scientology*).

Special cases for websites:

- reference to the access to the site: `<loc>` :
  Lesen sie den Artikel auf lemonde.fr;
  retrouvez cet article sur lemonde.fr

- reference to the site as a whole: `<prod.media>` :
  Interview auf lemonde.fr, mediapart.fr zeigt, dass Eric Woerth 50.000 Euro erhalten hat; Interview à retrouver sur lemonde.fr, mediapart.fr indique que Eric Woerth a bien touché 50.000 euros

- reference to the company that publishes the site: `<org>` :
  Sarkozy bemängelt mediapart.fr;
  Sarkozy dénonce mediapart.fr

Site addresses (www.radio-france.fr) are annotated as `<loc>`. However *Le site internet Radio France* is not an entity named in itself (we annotate only *Radio France* with `prod.media`).

3.5 Non-annotated entities

- Expressions of time (unlike in Impresso)
- Human productions (unlike in Impresso)
- Names of diseases (*AIDS, Grippe A; SIDA,* etc.)
- Psychological phenomena (*Ödipuskomplex; syndrome de Stockholm*, etc.)
- Scientific terms cannot be reduced to a product (*DNA, ADN,* etc.)
- Teaching programmes (*Staps, DEUG*, etc.)
- Special contracts (*le contrat Coca-Cola/Danone,* etc.)
  However: in *le contrat Coca-Cola*, the entity *Coca-Cola* is annotated (`org.ent`).

CC BY

- Political and/or judicial matters (*Watergate, Monica-gate; affaire Dickinson,* etc.).
  Optional: these may fall into a category depending on the assessments of the annotators.
- Climatic phenomena (*der Sturm Yinthya, le Mistral,* etc.).
  Optional: these may fall into a category depending on the assessments of the annotators.
- Social phenomena (*l'immigration arménienne* [11], etc.).
  Optional: these may fall into a category depending on the assessments of the annotators.

  NOTE: In some cases, it is still necessary to annotate the components of these expressions.
  - we do not annotate *Stockholm Syndrome* but we must annotate *Stockholm* (<loc>)
  - we do not annotate *complex d'Oedipus* but we must annotate *Oedipus* (<pers>)
  - we do not annotate *Statue of Pushkin* but we must annotate *Pushkin* (<pers>)

4. Quick guide and concrete considerations

4.1 Punctuation marks

All punctuation marks (including apostrophes) attached to named entities are left as separate tokens. They are not annotated except when they belong named entities such as for addresses, acronyms and abbreviations. Here are some examples:

(Berlin)
(<loc> Berlin </loc>)

Dr. Duboc
<per> Dr . Duboc </per>

H. C. Lausanne
<org>*H . C . Lausanne*</org>

Quai du Mont-Blanc, Geneva
<loc> Quai du Mont-Blanc, Geneva </loc>

Dr. Duboc lives in *Berlin.*
<per> Dr . Duboc </per> lives in <loc> Berlin </loc>.

4.2 Hesitations

A. Checking

If you need to double check a point, please use these resources:

- for German, Duden: http://duden.de
- for French, Larousse (tab 'Dictionary' or 'Encyclopedia'): https://www.larousse.fr/dictionnaires/francais

In case you suspect something to be a named entity but a quick check on the above mentioned resources and/or Wikipedia does not give information, skip the annotation.

[11] However, this term is annotated if it refers to a group of people rather than a process, see section 2.3.1.2.

CC BY

B. Reporting hesitations

For any dubious cases, please report you questions with screenshot and comments at the end of this file, ideally with screenshots including context, and annotation options:
https://docs.google.com/document/d/1yg7MGSfOvPnGoSXBWOuQaei0bY6xtTqtGTJtNuPyaeY/edit

C. Inception mini-tutorial

https://docs.google.com/document/d/1Tk6oadZNvVxHKSsLpVkPgOeKGs5kZTbgKxnsTpLB-vM/edit?usp=sharing

4.3 Overview of types, subtypes and components

Entity types and subtypes	
`pers`	• A single person (*Roger Federer*) • A named group of people including musical groups (die *Beatles*, La *Mano Negra*). (note: die *Schweizer*, Les *français* are **not** annotated.)
`pers.articleauthor`	A single person who is the author of an article.
`org`	• Organization that markets products or provides services (Die *Peugeot Gesellschaft,* Die *Waid;* La *société Peugeot*, la *Pitié-Salpêtrière*). (note: Die *schweizer* Polizei; la police francaise ist **not** annotated) • Including special type related to newspaper to spot press agencies (a subtype for Impresso v2.0).
`loc`	• District, locality, hamlet, village, city, etc. (*Paris*, *Val de Crüye*). • Cantons, communities of municipalities, departments, regions, etc. (*Autonome Gemeinschaft Baskenland;* les *Bouches du Rhône*, Le *Pays-Basque espagnol*). • Countries (*Schweiz*; *France*). • World regions, continent (*Maghreb; Pays-Basque*). • Mountains, plains, plateaus, caves, volcanoes, canyons (Die *Alpen,* Der *Vesuv; gouffre de Padirac*, Le *mont Ventoux*). • Oceans, seas, rivers, streams, ponds, marshes (Der *Atlantik,* Der *Golfstrom;* La *Seine*, Le *Lac Paladru*). • Planets, stars, galaxies and their parts (Der *Mond*, Die *Milchstrasse;* La *terre,* la *mer de la Tranquillité*) • Roads, highways, streets, avenues, squares, etc. (Die *Autobahn A6;* L'*autoroute A6*).

CC BY

	● Buildings (Der *Prime Tower;* Le *Palais de l'Élysée*). ● Physical addresses (*LIMSI-CNRS*, *Bâtiment 508, BP133, 91403 Orsay Cedex*). ● Electronic contact information (telephone and fax numbers, URL, e-mail address, identification of social network or Internet communication tools, etc., *http://www.limsi.fr/, 01-69-85-80-00*)
`prod.media`	Newspapers, magazines, broadcasts, sales catalogues, etc. (*Die Zeit;* Le *Figaro*, *Le sept à huit, La ferme célébrités*).

5. Stance annotation guidelines

Stance annotation consists of deciding whether an author of a text talking about an entity in a positive/favorable or in a negative/unfavorable light, or if the statement is rather objective/neutral. Three cases of stances can thus be distinguished: two cases of subjectivity, in which case we can directly indicate the polarity (POS, NEG), and the case of non-subjectivity, objectivity or neutrality (OBJ).
OBJ is the default option, so there is no need to label neutral/objective examples.

Since stance detection is a new task, we believe that the guidelines will be enriched alongside the annotation, ambiguities will be explored gradually as tests and annotations. The more examples we have, the better it is.

In order to define a starting standard to annotate stances toward topics and named entities in a piece of text, we propose below some suggestions and clarifications that may help.

1. We are not interested in knowing author's feeling but we look for author's stance with respect to a target entity. The stance expressed towards the entity is not related to whether the whole piece of text is positive or negative.
2. We have to separate good/bad news from the stance expressed. We should NOT annotate the good/bad content of the news. E.g. if the news talks about the damage of the fire of the Notre Dame Cathedral, the stance with respect to the Cathedral is objective (OBJ), even if this is considered bad news.
3. The annotator can imagine that he is the one being talked about: would he like or dislike the statement?
4. In case of doubt, it is absolutely recommended to not mark the stance. It will be considered as OBJ, the default option.
5. If an entity X indicates the faults of another entity Y in the text, note that the stance is negative only towards Y, the stance towards X is neutral.

CC BY

6. Named entity linking guidelines

Named Entity Linking (NEL) aims to disambiguate entities by linking them to entries of a Knowledge Base (KB). The following subsections provide some explanations about the annotation of named entity linking.

6.1 How Specific Should Linked Entities Be?

It is important to resolve disagreement when more than one annotation is plausible. The TAC-KBP annotation guidelines (tac, 2012) specify that different iterations of the same organization (e.g. the KB:111th U.S. Congress and the KB:112th U.S. Congress) should not be considered as distinct entities.

Example

Adams and Platt are both injured and will miss England's opening World Cup Qualifier against Moldova on Sunday. (AIDA)

Here the mention "World Cup" is labeled as KB:1998 FIFA World Cup, a specific occurrence of the event KB:FIFA World Cup. Therefore, the real entity is KB:FIFA World Cup.

6. 2 Metonymy

Another situation in which more than one annotation is plausible is metonymy, which is a way of referring to an entity not by its own name but rather a name of some other entity it is associated with.

Example

Moscow's as yet undisclosed proposals on Chechnya's political future have , meanwhile, been sent back to do the rounds of various government departments. (AIDA)

The mention here, "Moscow", could be labeled as KB:Government of Russia, KB:Moscow(the city) or KB:Russia. However, neither the city nor the country can actually make a proposal. The real entity in play is KB:Government of Russia.

CC BY

ANNEX A Main changes w.r.t Quaero v1

- reduction of the type of linguistic expressions considered as named entity (predominance of proper name)
- reduction if the components taken into account
- addition of 2 subtypes: pers.ind.artauthor and org.ent.pressagency
- the 2 subtypes of org.adm and org.ent are kept (w.r.t to quaero v2)

ANNEX B Main changes w.r.t Impresso v2

- New Preamble
- Removed NE types “human productions” and “time”, updated section “non-annotated entities” accordingly
- Minor additional changes in relation to the above

ANNEX C Main changes w.r.t NewsEye v3

- Removed most NE subtypes, except <pers.ind.articleauthor>, renamed <pers.articleauthor> and <prod.media>
- Briefly, our types and changes from v2 are the following:
    - <pers>: everything as in Impresso, except we ignore all subtypes (thus mark person NEs as <pers>) with one exception: <pers.articleauthor>
    - <org>: everything as in Impresso, except we ignore all subtypes (thus mark organisation NEs as <org>)
    - <loc>: everything as in Impresso, except we ignore all subtypes (thus mark organisation NEs as <loc>)
    - <loc>: everything as in Impresso, except we ignore all subtypes (thus mark organisation NEs as <loc>)
    - <prod>: we only use <prod.media>. This is our only type of <prod>.
- We removed everything related to components
- We added guidelines for the annotations of stance and for NEL

CC BY