A Multilingual Dataset for Named Entity Recognition, Entity Linking and Stance Detection in Historical Newspapers

Ahmed Hamdi, Elvys Linhares Pontes, Emanuela Boros, Thi Tuyet Hai Nguyen, Günter Hackl, Jose G. Moreno, Antoine Doucet

A Multilingual Dataset for Named Entity Recognition, Entity Linking and Stance Detection in Historical Newspapers

Ahmed Hamdi
ahmed.hamdi@univ-lr.fr
University of La Rochelle, L3i
La Rochelle, France

Elvys Linhares Pontes
elvys.linhares_pontes@univ-lr.fr
University of La Rochelle, L3i
La Rochelle, France

Emanuela Boros
emanuela.boros@univ-lr.fr
University of La Rochelle, L3i
La Rochelle, France

Thi Tuyet Hai Nguyen
hai.nguyen@univ-lr.fr
University of La Rochelle, L3i
La Rochelle, France

Günter Hackl
g.hackl@readcoop.eu
Innsbruck University Innovations GmbH
Innsbruck, Austria

Jose G. Moreno
jose.moreno@irit.fr
University of Toulouse, IRIT
Toulouse, France

Antoine Doucet
antoine.doucet@univ-lr.fr
University of La Rochelle, L3i
La Rochelle, France

ABSTRACT

Named entity processing over historical texts is more and more being used due to the massive documents and archives being stored in digital libraries. However, due to the poor annotated resources of historical nature, information extraction performances fall behind those on contemporary texts. In this paper, we introduce the development of the NewsEye resource, a multilingual dataset for named entity recognition and linking enriched with stances towards named entities. The dataset is comprised of diachronic historical newspaper material published between 1850 and 1950 in French, German, Finnish, and Swedish. Such historical resource is essential in the context of developing and evaluating named entity processing systems. It evenly allows enhancing the performances of existing approaches on historical documents which enables adequate and efficient semantic indexing of historical documents on digital cultural heritage collections.

CCS CONCEPTS

• **Information systems** → *Information retrieval*; **Digital libraries and archives**; • **General and reference** → **Cross-computing tools and techniques**.

KEYWORDS

datasets, multilingual, diachronic historical newspapers, named entity recognition, entity linking, stance detection

SIGIR '21, July 11–15, 2021, Virtual Event, Canada.

ACM ISBN 978-1-4503-8037-9/21/07…$15.00
https://doi.org/10.1145/3404835.3463255

ACM Reference Format:
Ahmed Hamdi, Elvys Linhares Pontes, Emanuela Boros, Thi Tuyet Hai Nguyen, Günter Hackl, Jose G. Moreno, and Antoine Doucet. 2021. A Multilingual Dataset for Named Entity Recognition, Entity Linking and Stance Detection in Historical Newspapers. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21), July 11–15, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3404835.3463255

1 INTRODUCTION

Information extraction (IE) from unstructured data may include tasks such as named entity recognition (NER), entity linking (EL), and stance detection (StD). NER consists of locating named entities and categorising them into a set of pre-defined classes (i.e. person, location, organisation, etc.), EL allows the disambiguation of the recognised named entities to a knowledge base (i.e. Wikipedia, Wikidata) while StD aims to determine whether the author's stance is subjective or neutral towards a target entity and if it is subjective, thus it determines its polar orientation (i.e. favour or against).

As the massive amount of digitised archival material is increasing during the last few decades, information extraction from historical newspapers has become highly required. Unlike contemporary data for which there are numerous information extraction resources, historical documents face not only a serious bottleneck concerning the lack of annotated data but also the unavailability of data in low-resourced languages. For named entity processing, many previous works used contemporary datasets to handle historical texts [19]. However, they showed that contemporary resources are not very 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. For this reason, building historical resources is essential for NER and EL systems to achieve good performances over documents of historical nature. These resources are used either as a learning base to train models or as a reference to evaluate these models. Building multilingual resources for

NER and EL is even more crucial since the linguistic characteristic of historical archives is not as anglophone as it is in the current situation [17].

Extracting entities from historical collections faces two additional challenges compared to contemporary texts. On one hand, historical data is noisy by nature as it is generated by optical character recognition (OCR) engines from digitised images. On the other hand, spelling conventions in historical texts have continuously acquired significant changes, in particular with named entities where many of them vanished and/or have non-corresponding mentions in knowledge bases [24].

In this context, we built a multilingual resource based on diachronic historical newspaper material. This dataset identifies and categorises named entities in four languages: German, French, Finnish, and Swedish. We perform a large-scale annotation of scanned texts transcribed by OCR from several historical articles of different newspapers for each language. The annotations of named entities are enriched by adding the corresponding Wikidata link if available as well as the author's stance. We also accomplish a practical analysis of our dataset in multitasking evaluations.

The remainder of this paper is organised as follows. In Section 2, we present and discuss a selection of existing datasets. Then, in Section 3, a detailed description of our dataset is presented. A set of state-of-the-art models to tackle named entity recognition, entity linking, and stance detection are described and evaluated in Section 4.Finally, Section 5 concludes this paper and hints at future work.

2 HISTORICAL DATASETS

Many previous works focused on building annotated corpora for named entity recognition and linking. Most of them are either monolingual [2, 11, 23] or domain-specific [1, 6, 22]. As we are interested by the news domain, which emerges as the best-resourced, we describe in this section the main publicly available corpora extracted from historical newspapers.

2.1 HIPE Dataset

The HIPE dataset[1] was created by the organisers of the CLEF 2020 Evaluation Lab HIPE challenge [8]. It is composed of articles from several Swiss, Luxembourgish, and American historical newspapers from 1790 to 2010 [9]. More concisely, the German articles were collected from 1790 to 1940, the French articles were extracted from issues published from 1790 to 2010 while the English articles were collected from 1790 to 1960. The corpus was manually annotated by native speakers following the *impresso* guidelines [10], derived from the *Quaero* annotation guide[2] [21]. For German and French, the HIPE dataset was divided into three sets for train, validation and test while the English part was small-sized and divided into only validation and test partitions. The tagset defines five named entity types (persons, locations, organisations, human products and time) and includes 23 fine-grained subtypes.

[1] https://github.com/impresso/CLEF-HIPE-2020/tree/master/data
[2] Quaero guidelines

2.2 Quaero Dataset

The Quaero Old Press Extended Named Entity corpus[3] [20] is the first corpus that aimed to support the task of historical NER for French. The corpus is a large annotated corpus consisting of articles of three French newspaper titles (*Le Temps*, *La Croix* and *Le Figaro*) from the end of the 19^{th} century. The corpus is annotated over its OCR transcription with a character error rate of about 95% and a word error rate of more than 60%.

2.3 The Czech Dataset

The Czech Historical Named Entity Corpus (v1.0)[4] [12] is extracted from newspaper articles published in the second half of the 19^{th} century. The corpus is lemmatised and named entities are annotated using six entity types (person, institutions, artifacts & objects, locations, time and ambiguous entities). The annotation was manually carried out by two native speaker annotators and only annotations that are consonant were maintained in its published version.

2.4 Europeana Dataset

Europeana NER corpora[5] [16] is an annotated corpus derived from the large Europeana newspaper collection[6], composed of more than 1,000 digitised titles in over 40 different languages, spanning from 17^{th} to 20^{th} century. The NER corpus comprises historical newspapers in Dutch, French and German, containing essentially material published in the 19^{th} century. The Europeana NER corpora were created by selecting pages with a minimum word-level accuracy of 80%. 100 pages were selected for French and Dutch while 200 pages were selected for German in order to consider also newspaper pages from Austria.

In this paper, we introduce the NewsEye dataset which, similarly to the HIPE and Europeana datasets, performs multiple European languages. However, unlike them, our dataset includes Finnish and Swedish languages. Moreover, the German and French corpora are extracted from historical newspapers different from those used in the HIPE and Europeana datasets. One more particularity to our dataset is that it is also useful for StD tasks rather than only for NER and/or EL. To the best of our knowledge, no StD gold standard ever addressed historical material.

3 THE NEWSEYE DATASET

The NewsEye dataset is collected through the National Library of France (BnF[7]), Austria (ONB[8]), and Finland (NLF[9]). It comprises four corpora: the French corpus is composed of items from digitised archives of nine newspapers (i.e. *L'Oeuvre*, *La Fronde*, *La Presse*, *Le Matin*, *Marie-Claire*,*Ce soir*, *Marianne*, *Paris Soir* and *Regards*) from 1854 to 1946. The German corpus contains articles extracted from four newspapers (i.e. *Arbeiter-Zeitung*, *Mittags-Zeitung*, *Illustrierte Kronen Zeitung* and *Neue freie Presse*) from 1864 to 1933. Finally,

[3] http://catalog.elra.info/en-us/repository/browse/ELRA-W0073/
[4] http://chnec.kiv.zcu.cz/
[5] https://github.com/EuropeanaNewspapers/ner-corpora
[6] https://www.europeana.eu/de/collections/topic/18-newspapers
[7] https://www.bnf.fr
[8] https://www.onb.ac.at
[9] https://www.kansalliskirjasto.fi

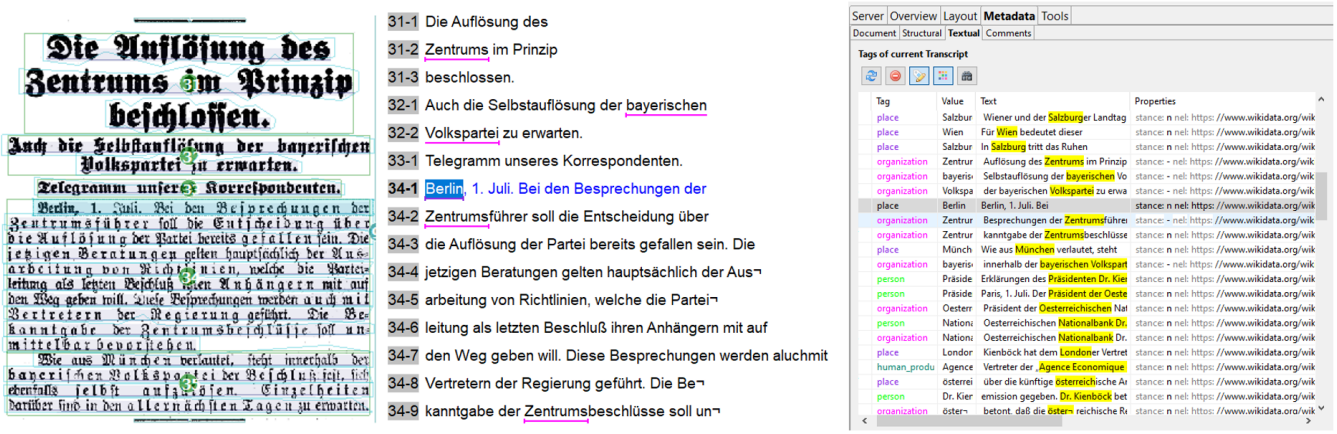

Figure 1: Annotation process using Transkribus, zoom on a newspaper article (left), its OCR output provided by the Austrian National Library (middle), and the annotation tool (right).

the Finnish and Swedish corpora, both comprise articles from two newspapers *Fraktur* and *Antique* published between 1852 and 1918 for Finnish and from 1848 to 1918 for Swedish.

3.1 Annotation Guidelines

The guidelines are established to define the named entity categories that would match the needs of the NewsEye users, on one hand, and to keep synergies with the Swiss-Luxembourg Impresso project, focused on building the CLEF-HIPE dataset (cf. Section 2.1), on the other hand. We, therefore, built the annotation guidelines in a concerted manner. The NewsEye annotation guidelines [10] actually started out as a branch of the Impresso guidelines [11] which are themselves derived from the Quaero annotation guide [12] utilised for building the Quaero dataset (cf. 2.2). Having annotations compatible across corpora would be beneficial for the community at large and in particular for the corresponding projects since datasets produced in one project could be used in the other projects. We believe that using similar guidelines across resources allow the community to take advantage of combined efforts with a significant amount of compatible data, rather than from independent and incompatible smaller collections.

Apart from a few fine-grained variations, the main difference with Impresso guidelines is that NewsEye guidelines focus on named entity 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.

We consider a named entity the real-world object denoting a unique individual with a proper name. We define four main types and one subtype of named entities in our dataset:

- person (PER): individual or group of persons;
  - authors of articles (PER.articleauthor) which indicate author 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 (HumanProd): we only focus on media products such as newspapers, magazines, broadcasts, etc.

As named entities can include one or more other named entities, our guidelines allow annotating nested named entities with a limit of depth one. Nested named entities are not considered in the EL and StD annotations.

Once NER guidelines are established, we defined the guidelines for entity linking and stance detection. The NER guidelines were detailed with a lot of examples to define named entities, their types, and their boundaries. To disambiguate all the difficult cases that can encounter the annotators, the guidelines for EL and StD were a bit more straightforward. They only provide some explanations about the annotation when more than one is plausible. For EL, two ambiguous cases were explained: metonymy and iterations. The metonymy consists of referring to an entity, not by its own name but rather a name of some other entity. The "White House", for instance, may refer to the location or to the US government and only the context can define which wikidata link should be assigned to the named entity in play. For iterations, we consider that all the occurrences of a named entity are not distinct (i.e. *the* 43^{nd} *and the* 44^{th} *of the ACM SIGIR International Conference* should refer to *ACM SIGIR International Conference*).

For the StD annotation, we provided some clarifications to annotators to specify the definition of a stance. As the task is rather new and not many guidelines are established in the literature, we proposed some suggestions and examples to distinguish between authors stances and author feelings (i.e. a good (resp. bad) news about a named entity does not necessarily mean that the stance with respect to this named entity is positive (resp. negative)).

[10] https://zenodo.org/record/4574199#.YD53r9zjKUk
[11] https://zenodo.org/record/3677171#.X8rwPLNCdPY
[12] http://www.quaero.org/media/files/bibliographie/quaero-guide-annotation-2011.pdf

Table 1: Distribution of annotations according to the named entity types in the NewsEye dataset.

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

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

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

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

Once the guidelines were compiled, the Transkribus [13] tool is adapted to allow named entity annotations and prepared datasets to be annotated, following up on technicalities. Transkribus is designed as a research infrastructure that enables users such as scholars, librarians, archivists, or researchers to carry out all steps of the digitisation, text recognition, and text augmentation workflow on their own. This also includes the creation of training data no matter if it is training for text, articles, or named entities. As part of the NewsEye project, the Transkribus platform was enhanced and augmented with new features and tools for processing historical newspapers. At the moment, over 50,000 Transkribus users can benefit from this development.

The main enhancement concerning named entities was the possibility to export them in the IOB format. There were also several technical features added to fasten the annotation of named entities, e.g. using shortcuts for the different categories or assigning properties - as in the Wikidata links - for several named entities at once with the help of a search table. After that, the annotators were trained to fulfill the task in a reasonable time.

In order to evaluate the IAA for the NewsEye dataset, several pages from each corpus have been annotated by two groups of native speakers of the concerned language. We then compute the IAA using the Kappa coefficient introduced by Cohen [5]. Table 1 shows the IAA for the NewsEye dataset and describes the distribution of annotations between the two groups. For each named entity type annotated by one group, we indicate how it was annotated by the other group.

Table 1 shows a 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 slightly less for organisations. In a few cases, named entities are associated with two different types by the two groups. This indicates that guidelines distinguish well the different types of named entities. 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.

As shown in Figure 2, the format used for annotation is CoNLL[14] where each word is annotated in a separate line. The TOKEN is the word from the full text; the TAG is the NER category (i.e. location and person in this example). The NESTED is annotated when the named entity appears inside another one. The SUBTYPE is limited to "author" when the named entity is a person. The WikiData-ID and STANCE are the fields respectively indicating the link to Wikidata for NEL and the stance value towards the concerned named entity. The IsSpaceAfter field is used specifically for languages that contain long compounded noun phrases (i.e. German) where sometimes only a part of a token is considered as a named entity. In that case, annotators split the NE part from the token and annotate them separately (in two lines), in this case, "NoSpaceAfter" is used to notify that the NE is a part of a token. The annotation follows the IOB scheme where O is used to annotated all tokens that are not named entities. B-label and I-label, respectively, indicate that the token is at the beginning (B-) of a NE or it is contained or at the end of it (I-).

```
TOKEN       TAG     NESTED  SUBTYPE WIKIDATA-ID STANCE  IsSpaceAfter

# -- Newspaper -- L_OEuvre / -- Issue -- 15-01-1930

COMEDIE     B-LOC   O       O       Q726531     n       SpaceAfter
DES         I-LOC   O       O       Q726531     n       SpaceAfter
CHAMPS      I-LOC   O       O       Q726531     n
-           I-LOC   O       O       Q726531     n
ELYSEES     I-LOC   O       O       Q726531     n
THEATRE     B-LOC   O       O       Q726531     n       SpaceAfter
LOUIS       I-LOC   B-PER   O       Q726531     n       SpaceAfter
JOUVET      I-LOC   I-PER   O       Q726531     n
AMPHITRTON  O       O       O       null        null    SpaceAfter
6U          O       O       O       null        null    SpaceAfter
THEATRE     B-LOC   O       O       Q3527480    n       SpaceAfter
DES         I-LOC   O       O       Q3527480    n       SpaceAfter
ARTS        I-LOC   O       O       Q3527480    n
A           O       O       O       null        null    SpaceAfter
LA          O       O       O       null        null    SpaceAfter
SCALA       B-LOC   O       O       Q3474920    n
```

Figure 2: Sample of the NewsEye dataset.

[13] https://transkribus.eu/

[14] https://universaldependencies.org/format.html

Table 2: Detailed description of the NewsEye dataset.

		Tags	Nested	Subtypes	Wiki-ID	Stance		
						Neutral	Negative	Positive
German	Person	3,500	246	30	685	3,384	59	57
	Location	5,904	544	–	610	5,840	15	49
	Organisation	3,370	176	–	967	3,334	33	3
	Human Production	44	2	–	17	43	0	1
French	Person	5,639	51	112	1,272	5,602	9	28
	Location	4,987	306	–	1,278	4,980	5	2
	Organisation	1,615	209	–	306	1,606	5	4
	Human Production	232	8	–	66	232	0	0
Finnish	Person	950	21	20	161	903	24	23
	Location	1,160	207	–	366	1,148	6	6
	Organisation	317	41	–	62	310	2	5
	Human Production	145	2	–	44	145	0	0
Swedish	Person	1,064	13	18	312	1,027	5	32
	Location	1,273	102	–	431	1,267	2	4
	Organisation	184	21	–	45	182	2	0
	Human Production	196	1	–	82	196	0	0

3.2 Statistics

The dataset contains 30,580 named entities with 6,704 Wikidata links. 214 positive stances and 167 negative were assigned to named entities. All the other entities were kept neutral with respect to the stance as to this extent it was to be expected for newspapers. Moreover, 180 persons were identified as authors of articles in this dataset.

The NewsEye dataset is segmented into sentences using the text-to-sentence splitter[15] which supports several European languages among them we can find German, French, Finnish, and Swedish. Each corpus is split into ∼ 80% for training (TRAIN), ∼ 10% for validation (DEV) and ∼ 10% for testing (TEST). The split strategy was made in a way that each newspaper issue takes part in each set. In addition, we also took care in our strategy to have well-distributed named entity types in each set with particular attention to non-frequent ones such as the *HumanProduct* (*HumanProd*) and the subtype *PER.author*. At the same time, the split was performed so that the positive and negative stances, which are dominated by the neutral stance (cf. Table 2), can be found in each partition with a reasonable balance. Table 3 describes the distribution of named entity types for each partition.

Table 3: Statistical description of the NewsEye partitions.

	Partition	Tokens	Entities	PER	LOC	ORG	HumanProd
German	Train	448,243	11,397	3,106	5,144	3,110	37
	Dev	40,062	539	149	263	123	4
	Test	39,451	882	245	497	137	3
	Total	527,756	12,818	3,500	5,904	3,370	44
French	Train	255,165	10,423	4,883	4,055	1,285	200
	Dev	21,727	752	293	335	113	11
	Test	30,458	1,298	463	597	217	21
	Total	307,350	12,473	5,639	4,987	1,615	232
Finnish	Train	48,222	2,146	782	979	259	126
	Dev	6,350	223	77	97	37	12
	Test	4,704	203	91	84	21	7
	Total	59,276	2,572	950	1,160	317	145
Swedish	Train	56,306	2140	838	985	153	164
	Dev	6,906	266	84	148	17	17
	Test	6,986	311	142	140	14	15
	Total	70,198	2,717	1,064	1,273	184	196

4 MODELS & EXPERIMENTS

The evaluation of the NER and EL is done in a coarse-grained manner, with the entity (not token) as the unit of reference [14]. We compute precision (P), recall (R), and F1 measure (F1) at the micro-level, i.e. error types are considered over all documents.

Due to the significant imbalance in the number of neutral and subjective stances, the stance detection model can obtain a very high F1-micro by predicting all stances as neutral. Therefore, F1-macro is reported instead of F1-micro in this task.

[15] https://github.com/mediacloud/sentence-splitter

4.1 Named Entity Recognition

We based our NER model on the best performing model at the CLEF 2020 Evaluation Lab HIPE challenge presented in [3, 4]. This architecture has as basis the pre-trained model BERT proposed by [7]. First, we use a pre-trained encoder, and second, we stack n Transformer blocks on top, finalised with a conditional random field (CRF) prediction layer.

A Transformer block (encoder), as proposed in [25], is a deep learning architecture based on multi-head attention. We refer to this model as BERT+n×Transformer where n is a parameter referring

to the number of Transformer layers.[16]. Table 4 shows the performance of the NER models for all languages on the test datasets.

Table 4: Results on NewsEye v1.0 with the best performing system `BERT-2XTransformer-CRF` for German and Swedish, and `BERT-1XTransformer-CRF` for French and Finnish.

Language	Precision	Recall	F1-micro
German	0.656	0.382	0.483
French	0.750	0.706	0.727
Finnish	0.785	0.770	0.777
Swedish	0.810	0.820	0.815

The additional Transformer layers proved that they can alleviate the sensitivity of the model towards out-of-vocabulary (OOV), OCR errors, or misspellings, and contributed to the learning and finding the proper informative words around entities [3].

4.2 Entity Linking

The disambiguation of entities in historical documents is a challenge because these documents contain OCR errors, spelling variations (language changes of old documents), and multilingualism (words in different languages).

In order to overcome these problems, we utilised the multilingual end-to-end entity linking (MEL) models described in [18] to process historical documents and disambiguate entities in Finnish, French, German, and Swedish. This system achieved the best results in terms of EL in the CLEF 2020 Evaluation Lab HIPE challenge [8]. To minimise the impact of historical documents on the EL task, this system is composed of modules to overcome problems related to multilingualism and OCR errors. The combination of probability tables of several languages provided different surface names for an entity in different languages.

Table 5 shows the performance of EL models for all languages on the test datasets. Compared to contemporary datasets (e.g. AIDA-CONLL dataset), EL systems achieved around $94 - 95$ points in precision [15]. This gap in the performance shows how historical data still a challenge for the EL task. The development of tools adapted to historical data (training data for deep learning and tools to minimise the OCR errors and language changes) can contribute to the improvement of EL systems.

Table 5: Results on NewsEye v1.0 with the best performing system for the entity linking task.

Language	Precision	Recall	F1-micro
German	0.725	0.472	0.571
French	0.750	0.536	0.625
Finnish	0.891	0.353	0.506
Swedish	0.750	0.436	0.552

[16]In our experiments, we have different values for n depending on their performance on the development set.

4.3 Stance Detection

The BERT-based models [7] showed an ability to process historical digitised data extracted from newspaper articles in many natural language processing (NLP) tasks. We consider the stance analysis toward given named entities as a sequence pair classification task. The first sequence is the body text, the second one is the given named entity, and the class label consists of positive, negative, or neutral values.

Table 6: Results on NewsEye v1.0 with the best performing system for the stance detection task.

Language	Precision	Recall	F1-macro
German	0.637	0.549	0.579
French	0.332	0.331	0.331
Finnish	0.434	0.497	0.460
Swedish	0.327	0.333	0.330

Similar to other fine-tuned models, the two sequences are tokenised by WordPiece [26] or SentencePiece [13]. Next, they are packed together into a single pair of sequences along with special classification tokens (i.e. *[CLS]* at the beginning and *[SEP]* at the end of each sequence). In addition, our approach randomly removes some neutral stances from the training data for alleviating the impact of imbalanced class distribution. Table 6 shows the performance of the stance detection models for all languages on the test datasets.

5 CONCLUSIONS AND FUTURE WORK

In this paper, we introduced the NewsEye dataset v1.0, a multilingual collection of annotated texts for historical named entity recognition, entity linking, and stance detection which is made publicly available for research purposes[17]. The NewsEye collection was extracted from several old press dispatches from different newspapers published from the middle of the 19^{th} century to the middle of the 20^{th} century and includes four European languages: French, German, Finnish, and Swedish. The annotation guidelines define four named entity types: person, location, organisation, and human product, and it is enriched with Wikidata links and authors' stances. The full annotation guidelines[18] are also uploaded and made available for the research community. The annotation reaches high inter-annotator agreements exceeding 0.8 Cohen's kappa for Swedish and 0.9 for German, French, and Swedish. We presented our most performing models along with the experimental results for establishing the baselines for all three tasks.

As future work, we plan to integrate more textual material, to release additional annotations. We also plan to detail the guidelines with respect to entity linking and stance detection which are being limited to give explanations and suggestions.

ACKNOWLEDGMENTS

This work has been supported by the European Union's Horizon 2020 research and innovation programme under grant 770299 (NewsEye).

[17]https://zenodo.org/record/4573313#.YD5FIdzjKUk
[18]https://zenodo.org/record/4574199#.YD53r9zjKUk

REFERENCES

[1] Sajawel Ahmed, Manuel Stoeckel, Christine Driller, Adrian Pachzelt, and Alexander Mehler. 2019. BIOfid Dataset: Publishing a German Gold Standard for Named Entity Recognition in Historical Biodiversity Literature. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. Association for Computational Linguistics, Hong Kong, China, 871–880. https://doi.org/10.18653/v1/K19-1081

[2] David Bamman, Sejal Popat, and Sheng Shen. 2019. An Annotated Dataset of Literary Entities. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 2138–2144. https://doi.org/10.18653/v1/n19-1220

[3] Emanuela Boroş, Ahmed Hamdi, Elvys Linhares Pontes, Luis-Adrián Cabrera-Diego, José G Moreno, Nicolas Sidere, and Antoine Doucet. 2020. Alleviating Digitization Errors in Named Entity Recognition for Historical Documents. In *Proceedings of the 24th Conference on Computational Natural Language Learning*. 431–441.

[4] Emanuela Boros, Elvys Linhares Pontes, Luis Adrián Cabrera-Diego, Ahmed Hamdi, José Moreno, Nicolas Sidère, and Antoine Doucet. 2020. Robust named entity recognition and linking on historical multilingual documents. In *Conference and Labs of the Evaluation Forum (CLEF 2020)*, Vol. 2696. CEUR-WS Working Notes, 1–17.

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

[6] Giovanni Colavizza and Matteo Romanello. 2017. Annotated References in the Historiography on Venice: 19th–21st Centuries. *Journal of Open Humanities Data* 3, 0 (Nov. 2017), 2. https://doi.org/10.5334/johd.9

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

[8] Maud Ehrmann, Matteo Romanello, Stefan Bircher, and Simon Clematide. 2020. Introducing the CLEF 2020 HIPE Shared Task: Named Entity Recognition and Linking on Historical Newspapers. In *European Conference on Information Retrieval*. Springer, 524–532.

[9] Maud Ehrmann, Matteo Romanello, Simon Clematide, Phillip Benjamin Ströbel, and Raphaël Barman. 2020. Language Resources for Historical Newspapers: the Impresso Collection. In *Proceedings of The 12th Language Resources and Evaluation Conference*. 958–968.

[10] Maud Ehrmann, Matteo Romanello, Alex Flückiger, and Simon Clematide. 2020. *Impresso Named Entity Annotation Guidelines*. Annotation Guidelines. Ecole Polytechnique Fédérale de Lausanne (EPFL) and Zurich University (UZH). https://doi.org/10.5281/zenodo.3604227

[11] Sara Grilo, Márcia Bolrinha, João Silva, Rui Vaz, and António Branco. 2020. The BDCamões Collection of Portuguese Literary Documents: A Research Resource for Digital Humanities and Language Technology. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 849–854. https://www.aclweb.org/anthology/2020.lrec-1.106

[12] Helena Hubková, Pavel Kral, and Eva Pettersson. 2020. Czech Historical Named Entity Corpus v 1.0. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 4458–4465. https://www.aclweb.org/anthology/2020.lrec-1.549

[13] Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Brussels, Belgium, 66–71. https://doi.org/10.18653/v1/D18-2012

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

[15] Isaiah Onando Mulang', Kuldeep Singh, Chaitali Prabhu, Abhishek Nadgeri, Johannes Hoffart, and Jens Lehmann. 2020. Evaluating the Impact of Knowledge Graph Context on Entity Disambiguation Models. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (Virtual Event, Ireland) *(CIKM '20)*. Association for Computing Machinery, New York, NY, USA, 2157–2160. https://doi.org/10.1145/3340531.3412159

[16] Clemens Neudecker. 2016. An Open Corpus for Named Entity Recognition in Historic Newspapers. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA), Portorož, Slovenia, 4348–4352.

[17] Clemens Neudecker and Apostolos Antonacopoulos. 2016. Making Europe's Historical Newspapers Searchable. In *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*. IEEE, 405–410.

[18] Elvys Linhares Pontes, Luis Adrián Cabrera-Diego, Jose G. Moreno, Emanuela Boros, Ahmed Hamdi, Nicolas Sidère, Mickaël Coustaty, and Antoine Doucet. 2020. Entity Linking for Historical Documents: Challenges and Solutions. In *Digital Libraries at Times of Massive Societal Transition*, Emi Ishita, Natalie Lee San Pang, and Lihong Zhou (Eds.). Springer International Publishing, Cham, 215–231.

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

[20] Sophie Rosset, Grouin, Cyril, Fort, Karen, Galibert, Olivier, Kahn, Juliette, and Zweigenbaum, Pierre. 2012. Structured Named Entities in Two Distinct Press Corpora: Contemporary Broadcast News and Old Newspapers. In *6th Linguistics Annotation Workshop (The LAW VI)*. Association for Computational Linguistics, Jeju, South Korea, 40–48.

[21] Sophie Rosset, Grouin, Cyril, and Zweigenbaum, Pierre. 2011. *Entités Nommées Structurées : Guide d'annotation Quaero*. Technical Report 2011-04. LIMSI-CNRS, Orsay, France.

[22] Rachele Sprugnoli. 2018. Arretium or Arezzo? A Neural Approach to the Identification of Place Names in Historical Texts. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-It 2018) (CEUR Workshop Proceedings, Vol. 2253)*, Elena Cabrio, Alessandro Mazzei, and Fabio Tamburini (Eds.). CEUR-WS, Torino, Italy, 1–6. http://ceur-ws.org/Vol-2253/paper26.pdf

[23] Sara Tonelli, Rachele Sprugnoli, and Giovanni Moretti. 2019. Prendo La Parola in Questo Consesso Mondiale: A Multi-Genre 20th Century Corpus in the Political Domain. In *Proceedings of the Sixth Italian Conference on Computational Linguistics, Bari, Italy, November 13-15, 2019 (CEUR Workshop Proceedings, Vol. 2481)*, Raffaella Bernardi, Roberto Navigli, and Giovanni Semeraro (Eds.). CEUR-WS.org, Bari, Italy, 1–8. http://ceur-ws.org/Vol-2481/paper71.pdf

[24] Seth Van Hooland, Max De Wilde, Ruben Verborgh, Thomas Steiner, and Rik Van de Walle. 2015. Exploring entity recognition and disambiguation for cultural heritage collections. *Digital Scholarship in the Humanities* 30, 2 (2015), 262–279.

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

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