Named Entity Disambiguation and Linking on Historic Newspaper OCR with BERT

Kai Labusch[1] and Clemens Neudecker[1]

Staatsbibliothek zu Berlin -
Preußischer Kulturbesitz
10785 Berlin, Germany
{kai.labusch,clemens.neudecker}@sbb.spk-berlin.de

Abstract. In this paper, we propose a named entity disambiguation and linking (NED, NEL) system that consists of three components: (i) Lookup of possible candidates in an approximative nearest neighbour (ANN) index that stores BERT-embeddings. (ii) Evaluation of each candidate by comparison of text passages of Wikipedia performed by a purpose-trained BERT model. (iii) Final ranking of candidates on the basis of information gathered from previous steps. We participated in the CLEF 2020 HIPE NERC-COARSE and NEL-LIT tasks for German, French, and English. The CLEF HIPE 2020 results show that our NEL approach is competitive in terms of precision but has low recall performance due to insufficient knowledge base coverage of the test data.

Keywords: Named Entity Recognition· Entity Linking· BERT· OCR

1 Introduction

Our participation in the CLEF HIPE 2020 NER-COARSE and NEL-LIT task[1] has been conducted as part of the Qurator[2] project within the Berlin State Library (Staatsbibliothek zu Berlin - Preußischer Kulturbesitz, SBB). One goal of the SBB in the Qurator project is the development of a system that identifies persons, locations and organizations within digitized historical text material obtained by Optical Character Recognition (OCR) and then links recognized entities to their corresponding Wikidata-IDs. Here, we provide a high-level overview of the functionality of our system; for details, take a deeper look at the information provided together with the source code[3].

The paper is structured as follows: after a brief introduction of the background and use case, a short summary of the Named Entity Recognition system

[1] https://impresso.github.io/CLEFHIPE2020/
[2] https://qurator.ai
[3] https://github.com/qurator-spk/sbb_ned

is provided in chapter 2. Chapter 3 outlines the Entity Linking approach developed in greater detail. Chapter 4 covers the chosen method for evaluation of candidates for entity linking and chapter 5 continues with a description of their ranking. Following a discussion of the results obtained in the NER-COARSE and NEL-LIT tasks in chapter 6, we wrap up with some concluding remarks and potentials for further improvement in chapter 7.

1.1 Background

The SBB is continuously digitizing its copyright-free holdings and making them publicly available online in various formats for viewing and browsing[4] or automated[5] download. As part of an on-going process, a growing amount of OCR-derived full-texts of the digitized printed material is provided in ALTO[6] format for internal use cases such as full-text indexing and other information retrieval tasks.

With an increasing amount of digitized sources becoming available online, the need for automated ways of extracting additional information from these sources increases as well. Disciplines such as the Digital Humanities create use cases for text and data mining or the semantic enrichment of the full-texts with e.g. Named Entity Recognition and Linking (e.g. for the re-construction of historical social networks[7]).

The boost in popularity of neural networks in the early 2010s, which are not only capable of dealing with large amounts of data (i. e., big data), but also require enormous amounts of data to be trained on in order to produce high quality results, has addressed this need. However, due to the historical nature of the documents being digitized in libraries, standard methods and procedures from the NLP domain typically require additional adaptation in order to successfully deal with the historical spelling variation and the remaining noise resulting from OCR errors.

1.2 Qurator

The Qurator project[9], funded by the German Federal Ministry of Education and Research (BMBF), for a timeframe of three years (11/2018-10/2021), is based in the metropolitan region Berlin/Brandenburg. The consortium of ten project partners from research and industry combines vast expertise in areas such as Language as well as Knowledge Technologies, Artificial Intelligence and Machine Learning.

The project's main goal is the development of a sustainable technology platform that supports knowledge workers in various industries. The platform will

[4] https://digital.staatsbibliothek-berlin.de
[5] https://oai.sbb.berlin
[6] https://www.loc.gov/standards/alto/
[7] https://sonar.fh-potsdam.de/

simplify the curation of digital content and accelerate it dramatically. AI techniques are integrated into curation technologies and curation workflows in the form of domain specific solutions covering the entire life cycle of content curation. The solutions being developed focus on curation services for the domains of culture, media, health and industry.

Within the Qurator consortium, the SBB is responsible for the task are "Curation Technologies for Digitized Cultural Heritage". The main goals of this task area lie in the development and adaptation of novel, AI/ML-based approaches from the document analysis and NLP domains for the improvement of the quality of OCR full-texts and the semantic enrichment of the derived full-texts with NER and NEL. The baseline for this development are the digitized collections of the SBB, with approximately 175,000 (August 2020) digitized documents from the timeframe 1400–1920. While most of the documents are in German, there is great variation with many other European and also Asian languages being present in the collection. The collection comprises documents from a wide array of publication formats, including books, newspapers, journals, maps, letters, posters, and many more.

1.3 HIPE

The introduction of the CLEF HIPE 2020 shared task provided a welcome opportunity to assess the performance of our own NER and NED systems in comparison with others within the frame of a common and realistic benchmark setting. HIPE proposes two tasks, NER and NEL, for French, German and English, with OCRed historical newspapers as input. The SBB's digitization strategy has traditionally put a strong focus on historic newspapers, with projects like Europeana Newspapers[8] producing millions of pages of OCR from digitized newspapers.

Recent years have also brought about the application of deep learning models for NER and NEL, where HIPE first puts these developments to the test for more challenging historical and noisy materials. We therefore expect that many valuable insights and directions for future work will results from participation in the HIPE shared task.

2 Named Entity Recognition

Before entity disambiguation starts, the input text is run through a named entity recognition (NER) system that tags all person (PER), location (LOC) and organization (ORG) entities. For the CLEF HIPE 2020 task, we used a BERT[3] based NER-system that has been developed previously at SBB and described in [5].

We employed our off-the-shelf system[8] and did not use CLEF HIPE 2020 NER training data for fine-tuning. Our off-the-shelf system does not currently

[7] https://impresso.github.io/CLEF-HIPE-2020/
[8] https://github.com/qurator-spk/sbb_ner

support product (PROD) and time (TIME) entities. The German NER system has been trained simultaneously on recent and historical German NER ground truth. In case of French and English, we used our multilingual model, i.e., a single BERT model that was trained for NER on combined German, French, Dutch and English NER labeled data.

Starting from multilingual BERT-Base Cased, we applied unsupervised pre-training composed of the "Masked-LM" and "Next Sentence Prediction" tasks proposed by [3] using 2,333,647 pages of unlabeled historical German text from the DC-SBB dataset [6]. Furterhmore, we performed supervised pre-training on NER ground truth using the Europeana Newspapers [7], ConLL-2003 [12] and GermEval-2014 [1] datasets.

In the according cross-evaluation, it was found that unsupervised pre-training on DC-SBB data worsens BERT performance in the case of contemporary training/test pairs while the performance improves for most experiments that test on historical ground truth. The best performance for our model is achieved by combining pre-training using DC-SBB + GermEval + CoNLL and results obtained from that are comparable to the state-of-the-art (see table 1). For the discussion of the performance of our NER system in the particular context of HIPE, please see chapter 6.

		P	R	F_1
[5]	DC-SBB+GermEval+CoNLL	81.1 ±1.2	87.8 ±1.4	**84.3 ±1.1**
[10]	Newspaper (1703-1875)	-	-	**85.31**
[11]	Newspaper (1888-1945)	-	-	**77.51**

Table 1. Performance comparison of different historical German NER BERT models. Results in [5] were obtained by 5-fold cross validation, results in [10] and [11] have been obtained for some 80/20 training/test split.

3 Entity Linking: Lookup of Candidates

3.1 Construction of knowledge base for PER, LOC and ORG

Our entity linking and disambiguation works by comparison of continuous text snippets where the entities in question are mentioned. A purpose-trained BERT model (the evaluation model) performs that text comparison task (see chapter 4). Therefore, a knowledge base that contains structured information like Wikidata is not sufficient. Instead we need additional continuous text where the entities that are part of the knowledge base are discussed, mentioned and referenced. Hence, we derive the knowledge base such that each entity in it has a corresponding Wikipedia page since the Wikipedia articles contain continuous

Lang	PER	LOC	ORG	coverage of test data
DE	671398	374048	136044	71%
FR	217383	155856	39305	68%
EN	324607	198570	58730	47%

Table 2. Size of knowledge-base per category per language. French and English knowledge bases are significantly smaller than the German one due to loss of entities in the Wikipedia - Wikidata mapping. Coverage of CLEF HIPE 2020 NEL-LIT test data Q-IDs is similar for German and French while being significantly worse for English.

text that have been annotated by human authors with references that can serve as ground truth.

The knowledge base has been directly derived from Wikipedia through the identification of persons, locations and organizations within the German Wikipedia by recursive traversal of its category structure:

- **PER:** All pages that are part of the categories "Frau" or "Mann" or of one of the reachable sub-categories of "Frau" and "Mann". One problem with this approach is that fictional "persons" are typically not contained in that selection.
- **LOC:** All pages that are part of the category "Geographisches Objekt" or one of its sub-categories. We exclude everything that is part of "Geographischer Begriff" or one of its sub-categories.
- **ORG:** All pages that are part of the category "Organisation" or one of its sub-categories.

Note: we plan to use the structured information of Wikidata in order to more reliably identify PER, LOC and ORG entities within Wikipedia which should make the heuristic approach of knowledge base creation obsolete.

Some pages might end up in multiple entity classes at the same time due to the category structure of the German Wikipedia. In order to create disjunct entity classes, we first remove from the entity class ORG everything that is also included in PER or LOC. In a second step, we remove everything from the entity class LOC that is also part of PER or ORG. It has been pointed out by one of our reviewers that this step is conceptually not required by our approach and it actually will be obsolete as soon as we identify PER, LOC and ORG entities on the basis of Wikidata.

To construct knowledge bases for French and English, we first map the identified German Wikipedia entity pages to their corresponding Wikidata-IDs and then the Wikidata-IDs back to the corresponding French and English Wikipedia pages. Table 2 shows the size of the knowledge bases per category and language. Note, that the knowledge bases for French and English are significantly smaller than the German one due to loss of many entities within the Wikipedia - Wikidata mapping.

What is the cause of that loss? We checked at random a number of entities of all types (PER,LOC,ORG) that had been lost in the mapping between the German and the French or English Wikipedia. In all cases Wikidata actually did not contain a reference to some English or French version of Wikipedia. Hence either there actually is not a French or English version of that Wikipedia page available or the correct linking has not been established so far. We expect to end up with much larger knowledge bases by use of structured data from Wikidata for identification of entities.

After the unmasked CLEF HIPE 2020 test data had been published, we computed the coverage of our per language knowledge bases, i.e., the percentage of Wikidata entity IDs (NEL-LIT) in the test data that actually can be found in the corresponding knowledge base. That percentage is an upper bound on the systems performance. As you can see in table 2, the coverage is similar for German and French (roughly 70%) whereas it is significantly worse for English (roughly 50%).

3.2 Entity Lookup Index

After the knowledge bases have been established, for each of them an entity lookup index is created by computation of BERT embeddings of the page titles of the identified PER, LOC and ORG Wikipedia pages. The BERT embeddings are obtained from a combination of different layers of the evaluation model (see chapter 4). The embedding vectors of the tokens of the page titles are stored in an approximative nearest neighbour (ANN) index [2]. We use cosine similarity as distance measure and the ANN index uses 100 random projection search trees. There are separate ANN indices per supported language and per supported entity category.

Given some NER-tagged surface that is part of the input text, up to 400 linking candidates below a cut-off distance of 0.1 are selected by lookup of the nearest neighbours of the surface's embedding within the approximative nearest neighbour index of the corresponding language and entity category.

According to our observations the performance of our system improves with a higher number of candidates considered. Of course there is some upper limit to that, however, more important is the computational complexity that grows with the number of linking candidates. We did not systematically evaluate the effect of the number of linking candidates but we used a number of linking candidates that is sufficiently high. Note that there is also an interaction with the cut-off distance since in many cases there are not as many as 400 nearest neighbours within a distance of less than 0.1.

4 Evaluation of Candidates

For each entity of the knowledge bases (see chapter 3.1) there are text passages in Wikipedia where some human Wikipedia editor has linked to that particular entity. How many linked text passages we have for some particular entity differs

SQL-table "sentences"	
id	: A unique number that identifies each sentence.
text	: A JSON-array that contains the tokens of the sentence. Example: ["Der", "Begriff", "wurde", "von", "Georg", "Christoph", "Lichtenberg", "eingebracht", "."]
entities	: A JSON-array of same length as "text" that contains for each token of the sentence the target entity if the token is part of an Wikipedia link that has been created by some Wikipedia author. If a token is not part of an Wikipedia link its corresponding entity is empty. Example: ["", "", "", "", "Georg_Christoph_Lichtenberg", "Georg_Christoph_Lichtenberg", "Georg_Christoph_Lichtenberg", "", ""]

SQL-table "links"	
id	: A unique number that identifies each Wikipedia entity reference.
sentence	: The sentence-id of the sentence where the reference occurs (sentences.id).
target	: The target entity of the reference. Example: "Georg_Christoph_Lichtenberg"

Table 3. The SQLITE sentence database consists of two tables. The "sentences" table contains all the sentences of the Wikipedia where some Wikipedia author referenced a PER, LOC, or ORG entity. The "links" table enumerates all references to PER, LOC or ORG entities in the Wikipedia. In order to get all the sentences of the German wikipedia where "Georg Christoph Lichtenberg" has been referenced by some Wikipedia author, the following SQL statement is used:
SELECT links.target, sentences.id, sentences.text, sentences.entities FROM links JOIN sentences ON links.sentence=sentences.id WHERE links.target=="Georg_Christoph_Lichtenberg"

widely depending on the entity. Some entities have thousands of links available whereas other entities have only very few.

We created a SQLITE database that provides quick access to the mentions of some particular entity. Using the Wikipedia page title of the entity as key, for instance, "Georg_Christoph_Lichtenberg", the database returns all sentences where some human editor explicitly linked to "Georg_Christoph_Lichtenberg". The database can be derived programmatically from the Wikipedia without any human annotation being involved. Table 3 gives a short description of the structure of the SQLITE database.

Using that database, we created a training dataset that consists of random sentence pairs (A,B) where sentences (A,B) either reference the same entity or different entities. That training dataset defines a binary classification problem: Do sentences A and B refer to the same item or not?

We trained a BERT model with respect to this binary classification problem per supported language that we call the "evaluation model" in the following. Given some arbitrary sentence pair (A,B), the evaluation model outputs the probability of the two sentences refering to the same item.

During entity disambiguation, we build up to 50 sentence pairs (A,B) for each candidate that has been found in the lookup step (see chapter 3.2). The sentence pairs are composed in such a way that sentence A is part of the input text where the entity that is to be linked is mentioned and sentence B is a

sentence from Wikipedia where that particular candidate has been linked to by a Wikipedia author. The higher the number of evaluated sentence pairs per candidate is, the more reliable the ranking model (see Section 5) can determine the overall matching probability. Again, the computational complexity increases with the number of sentence pairs. Additionally, in most cases there is only a very limited number of reference sentences from Wikipedia available such that it is not possible to generate a large number of unique sentence pairs. The choice of 50 sentence pairs is a trade-off that takes into account these considerations.

Application of the evaluation model to each sentence pair results in a corresponding matching probability. The sets of sentence pair matching probabilities of all candidates are then further processed by the ranking model (see Section 5).

5 Ranking of Candidates

During previous steps, sets of possible entity candidates have been obtained for all the parts of the input text that have been NER-tagged. For each candidate, a number of sentence pairs have been examined by the evaluation model, resulting in a set of sentence pair probabilities per candidate.

The ranking step finally determines an ordering of the candidates per linked entity according to the probability that it is the "correct" entity the part of the input text is actually referring to.

We compute statistical features of the sets of sentence pair probabilities of the candidates, among them: mean, median, min, max, standard deviation as well as various quantiles. Additionally we sort all the sentence pair probabilities and compute ranking statistics over all the candidates.

Then, based on the statistical features that describe the set of sentence pair probabilities of each candidate, a random forest model computes the overall probability that some particular candidate is actually the "correct" corresponding entity. The random forest model is the only component of our system where the CLEF HIPE 2020 data was used for training.

Finally the candidates are sorted according to the overall matching probabilities that have been estimated by the random forest model. The final output of our NED system is the sorted list of candidates where candidates that have a matching probability less than 0.2 are cut off.

Our NED system does not implement the NIL entity that means either it returns a non-empty list of Wikidata IDs that have been sorted in descending order according to their overall matching probabilities or the result is "-" if there is not any candidate that has matching probability above 0.2.

6 Results

Table 4 lists the NER performance of our off-the-shelf NER system (SBB) on the CLEF HIPE 2020 test data in the NER-COARSE-LIT task. Additionally, it

also contains the results of the best performing system (L3i). In case of the SBB system, strict NER performance is significantly worse than fuzzy NER performance. That observation holds for the L3i system too, however, for our system the effect is much more pronounced. Strict NER is a much more demanding task, nevertheless, we partly attribute the difference in performance to the training data of our NER system (see [5]), which has been created according to multiple slightly different NER annotation standards and also to the fact that we did not fine tune the NER system by using training data provided by the CLEF HIPE 2020 task organizers.

According to our observations, the OCR quality of the French data is slightly better than the German one and both French and German have better OCR quality than the English text material. By OCR quality, we primarily mean the overall quality of the entire text but not the mean Levenshtein distances of the entity text passages with respect to the original text. NER performance resembles that observation, i.e., French and German are comparable whereas English is significantly worse. Therefore our current hypothesis is that these differences are partly caused by the sensitivity of our NER-tagger to OCR noise within the surrounding text.

Lang	Team	Evaluation	Label	P	R	F_1
DE	L3i	NE-COARSE-LIT-micro-fuzzy	ALL	0.870	0.886	0.878
DE	SBB	NE-COARSE-LIT-micro-fuzzy	ALL	0.730	0.708	0.719
DE	L3i	NE-COARSE-LIT-micro-strict	ALL	0.790	0.805	0.797
DE	SBB	NE-COARSE-LIT-micro-strict	ALL	0.499	0.484	0.491
FR	L3i	NE-COARSE-LIT-micro-fuzzy	ALL	0.912	0.931	0.921
FR	SBB	NE-COARSE-LIT-micro-fuzzy	ALL	0.765	0.689	0.725
FR	L3i	NE-COARSE-LIT-micro-strict	ALL	0.831	0.849	0.840
FR	SBB	NE-COARSE-LIT-micro-strict	ALL	0.530	0.477	0.502
EN	L3i	NE-COARSE-LIT-micro-fuzzy	ALL	0.794	0.817	0.806
EN	SBB	NE-COARSE-LIT-micro-fuzzy	ALL	0.642	0.572	0.605
EN	L3i	NE-COARSE-LIT-micro-strict	ALL	0.623	0.641	0.632
EN	SBB	NE-COARSE-LIT-micro-strict	ALL	0.347	0.310	0.327

Table 4. NER-COARSE results of our (SBB) off-the-shelf BERT based NER system on the CLEF HIPE 2020 test data in comparison to the best performing system (L3i). The SBB system has not been trained on the CLEF HIPE 2020 data and does not support PROD and TIME entities. For German, the system has been trained on recent and historical German data simultaneously whereas for French and English, we employed a multilingual system that has been trained on German, Dutch, French and English data at the same time.

Table 5 shows the NEL performance of our system if our BERT based NER-tagging is used as input whereas table 6 contains the results that have been reported when the NER ground truth had been provided to the NEL system. The two tables show that NEL performance significantly improves if NER ground truth is provided.

Interestingly, the recall of the French and German NEL system is similar although the French knowledge base is significantly smaller than the German one. This observation can be explained by the fact that coverage of the test data of the knowledge bases for German in French is similar (see Table 2). We attribute the much lower recall for the English test data to much lower coverage of the test data of the English knowledge base (see Table 2).

Precision of the German and French SBB system is comparable, again precision of the English system is significantly worse, even if NER-groundtruth is provided. We explain in Section 5 that our system provides a list of candidates that have matching probability above 0.2 that is sorted in descending order according to the matching probability. Hence, given a bad coverage of the knowledge base, as it is the case for English, non matching candidates will inevitably move up in that sorted list, i.e., the drop in precision can also be explained by the bad coverage of the knowledge base.

Lang	Team	Evaluation	Label	P	R	F_1
DE	SBB	NEL-LIT-micro-fuzzy-@1	ALL	0.540	0.304	0.389
DE	SBB	NEL-LIT-micro-fuzzy-relaxed-@1	ALL	0.561	0.315	0.403
DE	SBB	NEL-LIT-micro-fuzzy-relaxed-@3	ALL	0.590	0.332	0.425
DE	SBB	NEL-LIT-micro-fuzzy-relaxed-@5	ALL	0.601	0.338	0.432
FR	SBB	NEL-LIT-micro-fuzzy-@1	ALL	0.594	0.310	0.407
FR	SBB	NEL-LIT-micro-fuzzy-relaxed-@1	ALL	0.616	0.321	0.422
FR	SBB	NEL-LIT-micro-fuzzy-relaxed-@3	ALL	0.624	0.325	0.428
FR	SBB	NEL-LIT-micro-fuzzy-relaxed-@5	ALL	0.629	0.328	0.431
EN	SBB	NEL-LIT-micro-fuzzy-@1	ALL	0.257	0.097	0.141
EN	SBB	NEL-LIT-micro-fuzzy-relaxed-@1	ALL	0.257	0.097	0.141
EN	SBB	NEL-LIT-micro-fuzzy-relaxed-@3	ALL	0.299	0.112	0.163
EN	SBB	NEL-LIT-micro-fuzzy-relaxed-@5	ALL	0.299	0.112	0.163

Table 5. NEL-LIT results with NER-tagging performed by our off-the-shelf system. French and German performance is similar, English is significantly worse. The stark performance differences between German and French versus English can mainly be explained by differences in coverage of the test data of the knowledge bases (see Table 2).

Table 7 reports on the best NEL-LIT results per team where the NER task has been performed by each teams own NER system. Table 8 reports on the best NEL-LIT results per team where the NER ground truth has been provided to

Lang	Team	Evaluation	Label	P	R	F_1
DE	SBB	NEL-LIT-micro-fuzzy-@1	ALL	0.615	0.349	0.445
DE	SBB	NEL-LIT-micro-fuzzy-relaxed-@1	ALL	0.636	0.361	0.461
DE	SBB	NEL-LIT-micro-fuzzy-relaxed-@3	ALL	0.673	0.382	0.488
DE	SBB	NEL-LIT-micro-fuzzy-relaxed-@5	ALL	0.686	0.389	0.497
FR	SBB	NEL-LIT-micro-fuzzy-@1	ALL	0.677	0.371	0.480
FR	SBB	NEL-LIT-micro-fuzzy-relaxed-@1	ALL	0.699	0.383	0.495
FR	SBB	NEL-LIT-micro-fuzzy-relaxed-@3	ALL	0.710	0.390	0.503
FR	SBB	NEL-LIT-micro-fuzzy-relaxed-@5	ALL	0.716	0.393	0.507
EN	SBB	NEL-LIT-micro-fuzzy-@1	ALL	0.344	0.119	0.177
EN	SBB	NEL-LIT-micro-fuzzy-relaxed-@1	ALL	0.344	0.119	0.177
EN	SBB	NEL-LIT-micro-fuzzy-relaxed-@3	ALL	0.390	0.135	0.200
EN	SBB	NEL-LIT-micro-fuzzy-relaxed-@5	ALL	0.390	0.135	0.200

Table 6. NEL-LIT results with NER ground truth provided. As expected, availability of NER ground truth significantly improves NEL results (see Table 5 for comparison). The stark performance differences between German and French versus English can mainly be explained by differences in coverage of the test data of the knowledge bases (see Table 2).

each team. In both tables, i.e., Table 7 and Table 8, the results have been sorted according to precision. It turns out that our SBB NEL-system performed quite competitively in terms of precision but rather abysmally in terms of recall.
We attribute the bad recall performance to multiple reasons:

- Due to the construction of the knowledge bases, many entities end up without representation. Even for German, that has the best coverage, coverage is only 71%.
- The lookup step of our NEL system has not been extensively optimized up to now. The embeddings, for instance, that are stored in the approximative nearest neighbour indices have been selected only on an initial guess basis and have not been optimized for performance. Which layers of the model to use and how to combine them heavily impacts the properties of the lookup step. Additionally the parameters of the approximative nearest neighbour indices such as type of similarity measure, number of lookup trees and cut-off distance, have been chosen on a initial guess basis too and could be further optimized.

7 Conclusion

The results of our participation in the HIPE task highlight where the biggest potential for improvement of our NER / NEL / NED system is to be expected:

Lang	Team	Evaluation	Label	P	R	F_1
DE	L3i	NEL-LIT-micro-fuzzy-relaxed-@5	ALL	0.627	0.636	0.632
DE	SBB	NEL-LIT-micro-fuzzy-relaxed-@5	ALL	0.601	0.338	0.432
DE	UvA.ILPS	NEL-LIT-micro-fuzzy-relaxed-@5	ALL	0.311	0.345	0.327
FR	L3i	NEL-LIT-micro-fuzzy-relaxed-@5	ALL	0.695	0.705	0.700
FR	SBB	NEL-LIT-micro-fuzzy-relaxed-@5	ALL	0.629	0.328	0.431
FR	IRISA	NEL-LIT-micro-fuzzy-relaxed-@5	ALL	0.560	0.490	0.523
FR	UvA.ILPS	NEL-LIT-micro-fuzzy-relaxed-@5	ALL	0.397	0.220	0.283
FR	ERTIM	NEL-LIT-micro-fuzzy-relaxed-@5	ALL	0.150	0.084	0.108
EN	L3i	NEL-LIT-micro-fuzzy-relaxed-@5	ALL	0.651	0.674	0.662
EN	UvA.ILPS	NEL-LIT-micro-fuzzy-relaxed-@5	ALL	0.304	0.458	0.366
EN	SBB	NEL-LIT-micro-fuzzy-relaxed-@5	ALL	0.299	0.112	0.163

Table 7. NEL-LIT results per team with NER-tagging performed by the teams own NER system. The results have been sorted according to precision.

Lang	Team	Evaluation	Label	P	R	F_1
DE	L3i	NEL-LIT-micro-fuzzy-relaxed-@5	ALL	0.696	0.696	0.696
DE	SBB	NEL-LIT-micro-fuzzy-relaxed-@5	ALL	0.686	0.389	0.497
DE	aidalight-baseline	NEL-LIT-micro-fuzzy-relaxed-@5	ALL	0.440	0.435	0.437
FR	L3i	NEL-LIT-micro-fuzzy-relaxed-@5	ALL	0.746	0.743	0.744
FR	SBB	NEL-LIT-micro-fuzzy-relaxed-@5	ALL	0.716	0.393	0.507
FR	Inria-DeLFT	NEL-LIT-micro-fuzzy-relaxed-@5	ALL	0.604	0.670	0.635
FR	IRISA	NEL-LIT-micro-fuzzy-relaxed-@5	ALL	0.590	0.588	0.589
FR	aidalight-baseline	NEL-LIT-micro-fuzzy-relaxed-@5	ALL	0.516	0.508	0.512
EN	L3i	NEL-LIT-micro-fuzzy-relaxed-@5	ALL	0.744	0.744	0.744
EN	Inria-DeLFT	NEL-LIT-micro-fuzzy-relaxed-@5	ALL	0.633	0.685	0.658
EN	UvA.ILPS	NEL-LIT-micro-fuzzy-relaxed-@5	ALL	0.607	0.580	0.593
EN	aidalight-baseline	NEL-LIT-micro-fuzzy-relaxed-@5	ALL	0.506	0.506	0.506
EN	SBB	NEL-LIT-micro-fuzzy-relaxed-@5	ALL	0.390	0.135	0.200

Table 8. NEL-LIT results per team with NER ground truth provided. The results have been sorted according to precision.

– OCR-performance is crucial since it is the start of the processing chain and OCR noise causes, as expected, bad results in all the subsequent processing steps.
– NEL recall performance of the SBB system has the biggest potential for improvement. An obvious path to improvement of recall performance is a better construction of the knowledge bases that should lead to a better overall representation of entities.
– An extensive evaluation and optimization of the lookup step that includes hardening against OCR noise could improve recall.
– NER results of other teams show that huge improvements in terms of NER performance even under the presence of noise are possible[4]. That improvement directly benefits the NED/NEL steps. We will therefore carefully evaluate how these improvements have been achieved in order to optimize our own NER-tagger.

Due to the diverse nature of the CLEF HIPE 2020 task data, in particular due to the differences in OCR quality, for us, the performance evaluation has resulted in valuable insights into our NER/NED/NEL system. The HIPE task data is in our opinion quite realistic, which means that we expect our system will have to handle similar data in the real world. Hence, we consider our participation in the HIPE competition as an important and constructive step on the path towards improving NER/NED processing of real world text material that has been obtained by OCR of historical documents.

References

1. Benikova, D., Biemann, C., Kisselew, M., Padó, S.: Germeval 2014 named entity recognition: Companion paper. Proceedings of the KONVENS GermEval Shared Task on Named Entity Recognition, Hildesheim, Germany pp. 104–112 (2014)
2. Bernhardsson, E.: Annoy: Approximate Nearest Neighbors in C++/Python (2018), https://github.com/spotify/annoy
3. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. CoRR **abs/1810.04805** (2018), http://arxiv.org/abs/1810.04805
4. Ehrmann, M., Romanello, M., Bircher, S., Clematide, S.: Introducing the CLEF 2020 HIPE shared task: Named entity recognition and linking on historical newspapers. In: Jose, J.M., Yilmaz, E., Magalhães, J., Castells, P., Ferro, N., Silva, M.J., Martins, F. (eds.) Advances in information retrieval. pp. 524–532. Springer International Publishing, Cham (2020)
5. Labusch, K., Neudecker, C., Zellhöfer, D.: BERT for Named Entity Recognition in Contemporary and Historic German. In: Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019): Long Papers. p. 1–9. German Society for Computational Linguistics & Language Technology, Erlangen, Germany (2019), https://corpora.linguistik.uni-erlangen.de/data/konvens/proceedings/papers/KONVENS2019_paper_4.pdf
6. Labusch, K., Zellhöfer, D.: OCR Fulltexts of the Digital Collections of the Berlin State Library (DC-SBB) (June 26th 2019), https://doi.org/10.5281/zenodo.3257041

7. Neudecker, C.: An open corpus for named entity recognition in historic newspapers. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). pp. 4348–4352. European Language Resources Association (ELRA), Portorož, Slovenia (May 2016), https://www.aclweb.org/anthology/L16-1689
8. Neudecker, C., Antonacopoulos, A.: Making europe's historical newspapers searchable. In: 2016 12th IAPR Workshop on Document Analysis Systems (DAS). pp. 405–410. IEEE, New York, NY, USA (April 2016), https://doi.org/10.1109/DAS.2016.83
9. Rehm, G., Bourgonje, P., Hegele, S., Kintzel, F., Schneider, J.M., Ostendorff, M., Zaczynska, K., Berger, A., Grill, S., Räuchle, S., Rauenbusch, J., Rutenburg, L., Schmidt, A., Wild, M., Hoffmann, H., Fink, J., Schulz, S., Seva, J., Quantz, J., Böttger, J., Matthey, J., Fricke, R., Thomsen, J., Paschke, A., Qundus, J.A., Hoppe, T., Karam, N., Weichhardt, F., Fillies, C., Neudecker, C., Gerber, M., Labusch, K., Rezanezhad, V., Schaefer, R., Zellhöfer, D., Siewert, D., Bunk, P., Pintscher, L., Aleynikova, E., Heine, F.: QURATOR: Innovative Technologies for Content and Data Curation. CoRR **abs/2004.12195** (2020), https://arxiv.org/abs/2004.12195
10. Riedl, M., Padó, S.: A named entity recognition shootout for German. In: Proceedings of ACL. pp. 120–125. Melbourne, Australia (2018), http://aclweb.org/anthology/P18-2020.pdf
11. Schweter, S., Baiter, J.: Towards robust named entity recognition for historic german. arXiv preprint arXiv:1906.07592 (2019), https://arxiv.org/abs/1906.07592
12. Tjong Kim Sang, E.F., De Meulder, F.: Introduction to the conll-2003 shared task: Language-independent named entity recognition. In: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4. pp. 142–147. CONLL '03, Association for Computational Linguistics, Stroudsburg, PA, USA (2003), https://doi.org/10.3115/1119176.1119195