

Entity Linking in Multilingual Newspapers and Classical Commentaries with BERT

Kai Labusch¹, Clemens Neudecker¹

¹Staatsbibliothek zu Berlin - Preußischer Kulturbesitz, 10785 Berlin, Germany

Abstract

Building on our BERT-based entity recognition and three stage entity linking (EL) system [1] that we evaluated in the CLEF HIPE 2020 challenge [2], we focused in the CLEF HIPE 2022 challenge [3] on the entity linking part by participation in the EL-only tasks. We submitted results for the multilingual newspaper challenge (MNC), the multilingual classical commentary challenge (MCC), and the global adaptation challenge (GAC). This working note presents the most important modifications of the entity linking system in comparison to the HIPE 2020 approach and the additional results that have been obtained on the **ajmc**, **hipe2020**, **newseye**, **topres19th**, and **sonar** datasets for German, French, and English. The results show that our entity linking approach can be applied to a broad range of text categories and qualities without heavy adaptation and reveals qualitative differences of the impact of hyperparameters on our system that need further investigation.

Keywords

entity-linking, BERT, historical multi-lingual newspaper, classical commentaries

1. Introduction

The Berlin State Library (Staatsbibliothek zu Berlin - Preußischer Kulturbesitz, SBB) is one of the largest research libraries in Germany. The SBB continuously expands its digital collections, which at the time of writing comprise of approx. 200,000 digitized works from roughly 1500 to 1950. A key component in providing online access to the digitized collections of SBB is a keyword search, that is based on the full text derived by application of Optical Character Recognition (OCR) to the scanned document pages.

Thanks to recent advances in OCR technology for historical documents due to deep learning [4, 5], the quality of the OCR results has now reached a level where Natural Language Processing (NLP) techniques like Named Entity Recognition (NER) and Entity Linking (EL) can be used to further enrich the unstructured texts and thereby enable new ways of searching and analyzing the digitized content.

In the QURATOR research project [6], the SBB had the opportunity to investigate the suitability of state-of-the-art NER [7] and EL [1] methods for digitized historical documents, which led to the SBB's participation in the CLEF HIPE 2020 Shared Task organized by Ehrmann et al. [2]. Since then, the EL system has been further improved and evaluated [8], with the CLEF HIPE 2022 EL-only Task providing a welcome opportunity to revisit the evaluation of the EL system


CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy

✉ Kai.Labusch@sbb.spk-berlin.de (K. Labusch); Clemens.Neudecker@sbb.spk-berlin.de (C. Neudecker)

🌐 <https://staatsbibliothek-berlin.de/> (K. Labusch); <https://staatsbibliothek-berlin.de/> (C. Neudecker)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

and determine what performance improvements the changes have brought in comparison with the previous results and those from other research groups, such as L3i, who obtained the best results for EL-only in CLEF HIPE 2020 [9] and were the only other participant in the EL-only task for CLEF HIPE 2022 [2]. The EL-only task focuses on entity linking while the required entity recognition information is provided based on ground-truth as additional input to the system [10]. This approach enables an independent evaluation of the EL system by avoiding the additional variation that is introduced due to named entity recognition differences or errors.

Also, with SBB being involved in multiple research projects, where different requirements and annotation standards for NER/EL are being used, the introduction of a new set of challenges related to the adaptation of systems for multilingual and multi-domain input data with differing annotation depth and quality in the training data, fits particularly well with the real world scenarios encountered at SBB.

Accordingly, SBB participated in the EL-only task where we submitted results for all three challenges, the Multilingual Newspaper Challenge (MNC), Multilingual Classical Commentaries Challenge (MCC) and Global Adaptation Challenge (GAC). The paper is structured as follows: Section 2 provides a brief description of our EL method, followed by a short explanation of the experimental setup in Section 3, results achieved in our CLEF HIPE 2022 participation are analyzed in more detail in Section 4 and Section 5 concludes the paper with a summary and outlook to future work.

2. Entity-Linking

The experiments have been performed using an improved version of the EL system¹ that has previously been evaluated in the CLEF HIPE 2020 Shared Task [1, 2]. Here, we only give a brief overview of that system and discuss the most relevant additions and modifications with respect to our CLEF HIPE 2020 participation. For a more detailed description of our EL system, we refer to our CLEF HIPE 2020 working notes [1].

2.1. Wikidata Knowledge Base

In the CLEF HIPE 2020 EL tasks, we learned that our knowledge base that was derived by means of the category structure of Wikipedia provided an insufficient coverage of the entities that are mentioned in the CLEF HIPE 2020 EL test data.

In order to construct a more comprehensive knowledge base, we introduced a knowledge base that is derived from Wikidata and Wikipedia. Wikidata is used in order to find persons, locations and organisations that have a corresponding Wikipedia article by means of a set of SPARQL-queries². Though the current knowledge base is limited to persons, locations, and organizations due to the SPARQL approach it is adaptable and can be augmented with additional entity types by addition of corresponding SPARQL-queries.

For all those relevant entities that have been found within Wikidata, all the sentences of any Wikipedia article where those entities have been linked by a Wikipedia author are extracted and

¹https://github.com/qurator-spk/sbb_ned/tree/79f3739dcb3b68ade798ab95c177f4bfb641ae52

²https://github.com/qurator-spk/sbb_tools/tree/d954888d10a80096f7be4d5e5b202ba782479300/sparql

made accessible in a database for use in the evaluation step. See Section 4 of [1] for a description of the database. Sentences from this database where the candidate match has been linked by a Wikipedia author are compared against sentences of the target text where the entity in question of is mentioned (see also Section 4 of [1]).

Our EL system can only identify those entities that have a corresponding Wikipedia page since it performs a number of text comparisons for each candidate match out of the knowledge base, i.e., it needs at least one reference in a Wikipedia article to compare that reference and its context against the occurrence in the target text.

Wikidata provides lots of additional information about the entities. For instance, we include date-of-birth and date-of-inception for persons resp. organisations in the knowledge base and utilize this information to exclude entities from the linking process based on the date of publication. Our EL system accepts a time constraint as input so that persons or organisations that according to Wikidata did not exist at that point in time are not considered in the entity linking process. During the challenge, we evaluated the "hipe2022:date" field and provided that date as time constraint to the EL system.

After the CLEF HIPE 2022 test data had been published, we computed the percentage of Wikidata-IDs in the test data that actually can be found in the knowledge base of our system. Table 1 shows the coverage per dataset and language. The coverage defines an upper limit for the performance of the EL system. It is in our case close to or even above 90% for all datasets except **ajmc** which is due to the fact that currently certain types of entities, e.g., work of art, are not contained in the knowledge base.

2.2. Lookup of Candidates

The first step of EL is the identification of entries of the knowledge base that possibly match with the text passage that has been tagged either as a person, location or organisation. The passage of text in question is termed target surface. Those entries that are selected from the knowledge base as possible matches are denoted as candidates and further processed in the second and third stages of the EL process. The maximum number of candidates considered is an important hyperparameter and denoted as max_{cand} in the following.

Our candidate selection uses an approximate nearest neighbour index where word embeddings of parts of the Wikipedia page titles of the entities of the knowledge base are stored. The parts of the page titles are obtained, first by application of regular expressions that replace or remove characters such as whitespace, punctuation and separator characters, and second by splitting the lower-cased result along remaining word boundaries, i.e., " ", "-", and "_". We plan to improve this procedure by utilization of the name variants of the entities provided by Wikidata instead of Wikipedia page titles.

Then, by use of word embeddings of parts of the target surface that are derived by the same algorithm, related entities can be determined in an approximate nearest neighbour search within the word embedding space. Another important hyperparameter is the cut-off-distance of the nearest neighbour search that is denoted with Δ_l . See also Section 3 of [1].

dataset	lang	coverage
ajmc	de	0.57
newseye	de	0.88
hipe2020	de	0.86
sonar	de	0.93
ajmc	fr	0.53
newseye	fr	0.93
hipe2020	fr	0.90
ajmc	en	0.54
hipe2020	en	0.93
topres19th	en	0.99

Table 1

Percentage of Wikidata-IDs of the test-data that are actually contained in our knowledge base that has been derived from Wikidata and Wikipedia.

2.3. Evaluation of Candidates

For each candidate that has been selected in the first step, a number of text comparisons is performed. From the sentence database (see Section 2.1), for each candidate the sentences of Wikipedia where it is referenced are retrieved and pairwise compared with the sentence that surrounds the target surface. The comparison is done by means of a BERT model that has been purpose trained on pairs of Wikipedia sentences to answer the binary question: "Do these two sentences refer to the same entity or not?". The outcome of the second stage is a set of matching probabilities of the sentence pairs of each candidate. See also Section 4 of [1].

2.4. Ranking of Candidates

In the third step for each candidate, a feature vector is computed that contains statistical descriptors, i.e., minimum, maximum, standard deviation, median, and 0.1, . . . , 0.9 quantiles of the sentence pair matching probabilities from the second step, together with additional features from the lookup step, such as same statistical descriptors of the word embedding similarities within the embedding space.

The per-candidate feature vectors are fed into a random forest model that outputs the overall probability that a particular candidate is actually the correct match for the target surface. The final output consists of all candidates sorted according to this matching probability. Δ_r , the ranking threshold, is another important hyperparameter. Candidates that have a matching probability below that cut-off probability will be discarded. See also Section 5 of [1].

3. Submission

The goal of our submission was to evaluate our EL system "as is" in order to determine the off-the-shelf performance for a variety of text categories and qualities. Since the ranking ML-models (random forest) that are in use at SBB have been trained on the CLEF HIPE 2020 EL ground truth [11], they are the only component that had to be re-trained to ensure that only permitted data

has been used for training. The BERT models that are used in the entity lookup and evaluation steps have been trained only on the digitized collections of the SBB and on Wikipedia sentences. The entries of the lookup index have been retrieved from Wikidata and Wikipedia as explained before. Therefore, we used those parts without any re-training or modification.

The ranking ML-models have been separately re-trained per language (de, fr, en) on the joined training sets of the **ajmc**, **hipe2020**, **newseye**, **topres19th**, and **sonar** datasets. The information about the dataset membership of a sample is not provided to the ranking models. The dev subsets have not been used for training.

For the submission two different configurations have been used:

- Run 1: maximum number of candidates (max_{cand}) = 25; lookup threshold (Δ_l) = 0.05; ranking threshold (Δ_r) = 0.2
- Run 2: maximum number of candidates (max_{cand}) = 50; lookup threshold (Δ_l) = 0.13; ranking threshold (Δ_r) = 0.2

In order to evaluate the impact of maximum number of candidates, lookup threshold and ranking threshold, we conducted additional experiments for both configurations, where we tried different ranking thresholds, e.g., $\Delta_r = 0.05, 0.1, \dots, 0.5$.

4. Results

All the plots are based on the NEL-LIT-micro-fuzzy-eval-TIME-ALL-LED-ALL-metric implemented in the CLEF HIPE 2022 scorer³. We do not repeat figures for other metrics, since the findings that are discussed in the following can also be qualitatively observed for metrics like macro-doc-strict, macro-doc-fuzzy or micro-doc-strict^{4,5}.

Figure 1 shows the F_1 -score, precision and recall obtained on the dev subsets of the datasets versus the corresponding measurements obtained on the test sets. For each combination of language and dataset, i.e., **ajmc**, **hipe2020**, **newseye**, **topres19th** and **sonar**, different parametrizations of the EL system have been evaluated (see Section 3). Pairs of measurements located below the diagonal indicate that performance on the dev subset is better than performance on the test subset, whereas pairs of measurements located above the diagonal mean that performance on the dev subset is worse than performance on the corresponding test subset.

The results show that the performance on the dev subsets is actually a suitable indicator for the performance on the test set. Hence parameter optimization on the dev-sets is feasible in order to optimize performance on the test sets. Results on the dev subset have a tendency to overestimate the performance. Comparison of precision and recall figures shows, that the cause is overestimation of both recall and precision and that it is most noticeable in case of the **newseye** and **sonar** datasets. Only in case of the French **ajmc** dataset the results on the test subset are clearly better than those on the dev subset.

The intra-dataset and inter-dataset variation of performance differs among the datasets and languages. Intra-dataset variation of results is higher for German than for French or

³<https://github.com/hipe-eval/HIPE-scorer>

⁴https://github.com/qurator-spk/sbb_ned/blob/3feacdd60807df2ce45f8d0430f04974cfc79919/notebook/HIPE-2022.ipynb

⁵<https://github.com/hipe-eval/HIPE-2022-eval>

Evaluation: NEL-LIT-micro-fuzzy-eval-TIME-ALL-LED-ALL

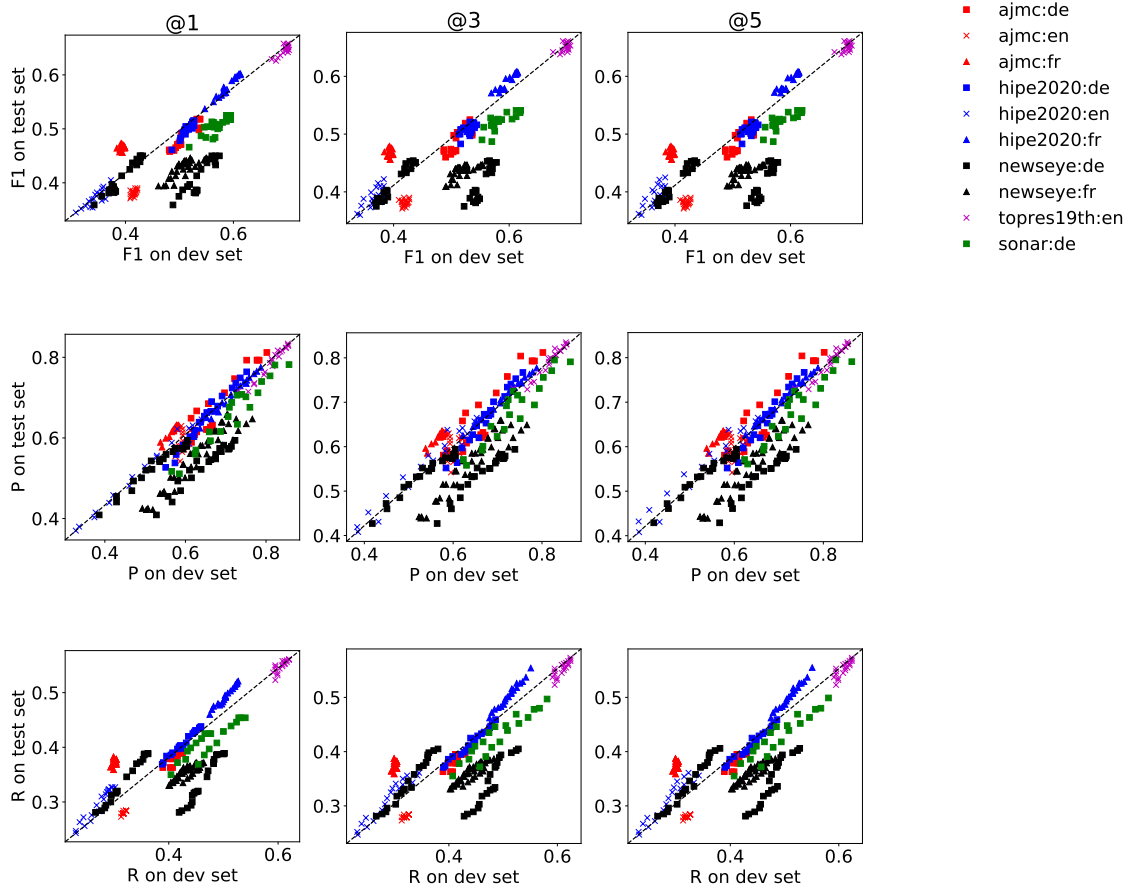


Figure 1: Observed F_1 score, precision and recall on the dev subsets versus the corresponding measurements on the test subsets for all combinations of max_{cands} , Δ_l , and Δ_r that have been evaluated. Color encodes the dataset whereas the shape of marker encodes the language.

English. For French and English, the measurements show distinct per-dataset clusters, while the measurements for German are much more spread out even within a particular dataset. We obtain both the best and the worst results for English, where performance depends mainly on the dataset.

Figure 2 depicts the results of the best performing hyperparameters on the dev subsets versus their corresponding measurements on the test subsets for each combination of dataset and language. Again, dev subset best-performance is a feasible indicator of test subset best-performance. The best-performance results are roughly symmetrically distributed around the diagonal. Table 2 lists the parameter combinations that have best dev subset performance, the corresponding test subset performance and for comparison the best-performing CLEF HIPE 2022 submission for each combination of dataset and language. All the dev subset optimized

Evaluation: NEL-LIT-micro-fuzzy-eval-TIME-ALL-LED-ALL

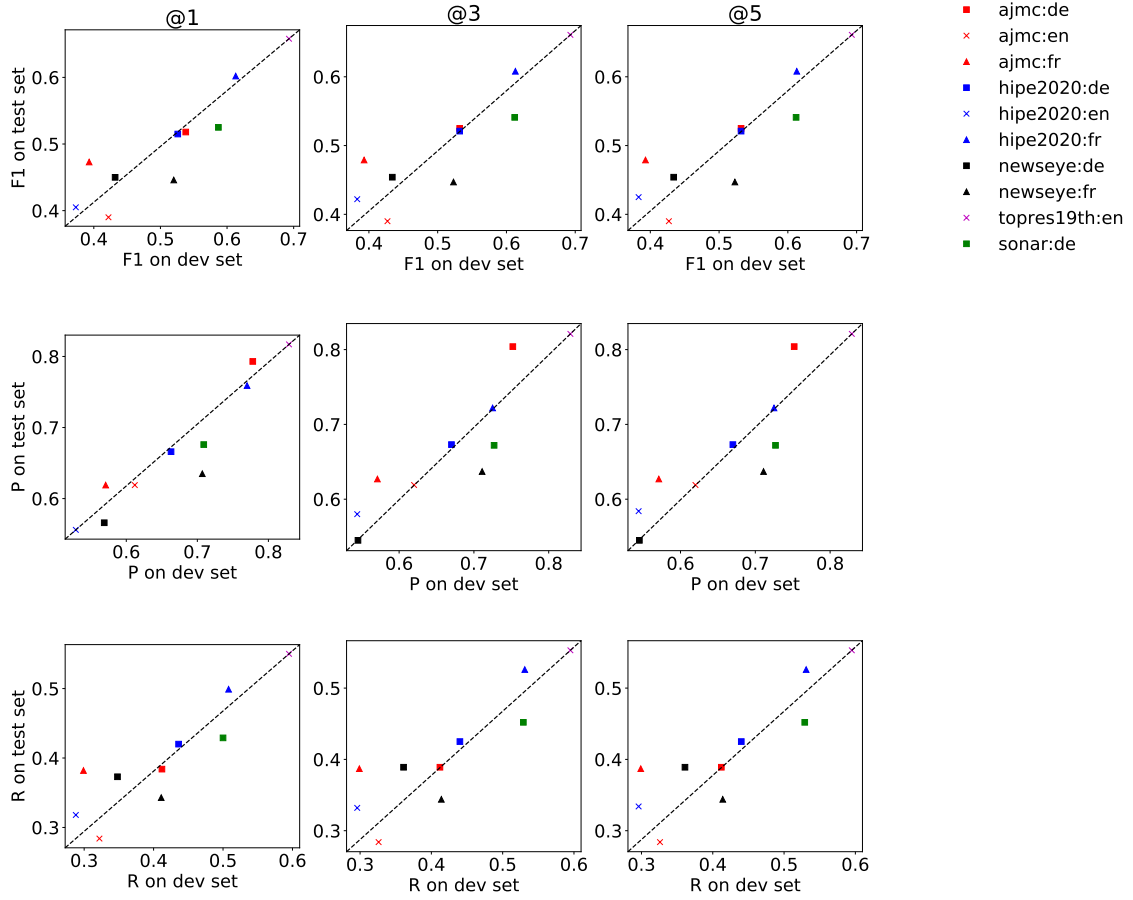


Figure 2: Observed F_1 score, precision and recall on the dev sets versus the corresponding measurements on the test set for the best combination of max_{cand} , Δ_l , and Δ_r per dataset and language that have been evaluated. Color encodes the dataset whereas the shape of marker encodes the language.

parameter combinations slightly outperform our best CLEF HIPE 2022 submission which were obtained with deliberately fixed off-the-shelf parameters. Overall comparison of these results with the coverage of the knowledge base in Table 1 leads us to the conclusion, that although the performance is weakly correlated to the coverage, it does not provide an explanation for the performance differences observed between different datasets and languages.

From Table 2 it can be seen, that a higher number of candidates for evaluation max_{cand} in combination with a larger lookup threshold Δ_l does not necessarily lead to best performance. Rather, it depends on the dataset and language if a higher number of evaluation candidates is beneficial for performance.

Figure 3 shows the impact of variation of the ranking threshold Δ_r on precision versus recall and its dependence on the number of evaluation candidates max_{cand} and lookup threshold

dataset	lang	Δ_l	Selection based on dev set				HIPE 2022 Submission			
			max_{cand}	Δ_r	$F_{1_{dev}}$	$F_{1_{test}}$	$F_{1_{sub}}$	Δ_r	$d\Delta_l$	max_{cand}
ajmc	de	0.13	50	0.35	0.538	0.518	0.503	0.2	0.13	50
ajmc	en	0.13	50	0.35	0.422	0.390	0.381	0.2	0.13	50
ajmc	fr	0.13	50	0.15	0.393	0.473	0.470	0.2	0.05	25
hipe2020	de	0.05	25	0.4	0.526	0.515	0.506	0.2	0.05	25
hipe2020	en	0.13	50	0.25	0.373	0.405	0.393	0.2	0.13	50
hipe2020	fr	0.13	50	0.4	0.613	0.602	0.596	0.2	0.13	50
newseye	de	0.05	25	0.35	0.432	0.450	0.444	0.2	0.05	25
newseye	fr	0.13	50	0.45	0.520	0.446	0.430	0.2	0.13	50
sonar	de	0.05	25	0.3	0.587	0.525	0.517	0.2	0.05	25
topres19th	en	0.13	50	0.35	0.693	0.658	0.651	0.2	0.13	50

Table 2

Best test performance in terms of F_1 -score by hyperparameter selection based on dev set performance versus best parameter configuration of our original CLEF HIPE 2022 submission. The evaluation metric is NEL-LIT-micro-fuzzy-eval-TIME-ALL-LED-ALL-@1.

Δ_l . For each combination of dataset, language, number of evaluation candidates and lookup threshold, one obtains a receiver-operator-characteristics-like trajectory of measurements by variation of the ranking threshold $\Delta_r = 0.05, \dots, 0.5$. The lower the ranking threshold, the better recall gets while precision decreases. Only in German, except for **ajmc**, it is clearly beneficial to use a lower number of evaluation candidates and a lower lookup threshold in high-recall and low-to-medium-precision regimes. In high-precision and low-recall regimes, a high number of evaluation candidates and a larger lookup threshold seems to be the better choice.

For French and English, we do not observe significant differences on the **ajmc**, **newseye**, and **topres19th** datasets, while for both languages on the **hipe2020** dataset, an improvement can be obtained by use of a higher number of evaluation candidates and a larger lookup threshold.

5. Conclusion

For an EL system that is meant to be applied to library collections of digitized text materials from a broad variety of domains, origins and qualities, covering a time span of over 400 years, it is first and foremost required to be robust and able to deliver a solid baseline performance - even when not optimized for a particular type of material, language or time frame. Furthermore, use cases for NER and EL at SBB derive from varying contexts such as retrieval, historical social network analysis or subject indexing, which have distinct requirements and application contexts each, demanding a system that can also easily be tailored to more specific use cases.

Our experiments performed in the CLEF HIPE 2022 challenges indicate that our EL system actually provides a robust baseline performance for a variety of text categories, qualities and origins, though the overall performance can still be greatly improved.

Comparison of our results for dev subsets versus test subsets shows consistent outcomes for all the datasets. Our improved Wikidata-based construction of the knowledge base provides

Evaluation: NEL-LIT-micro-fuzzy-eval-TIME-ALL-LED-ALL

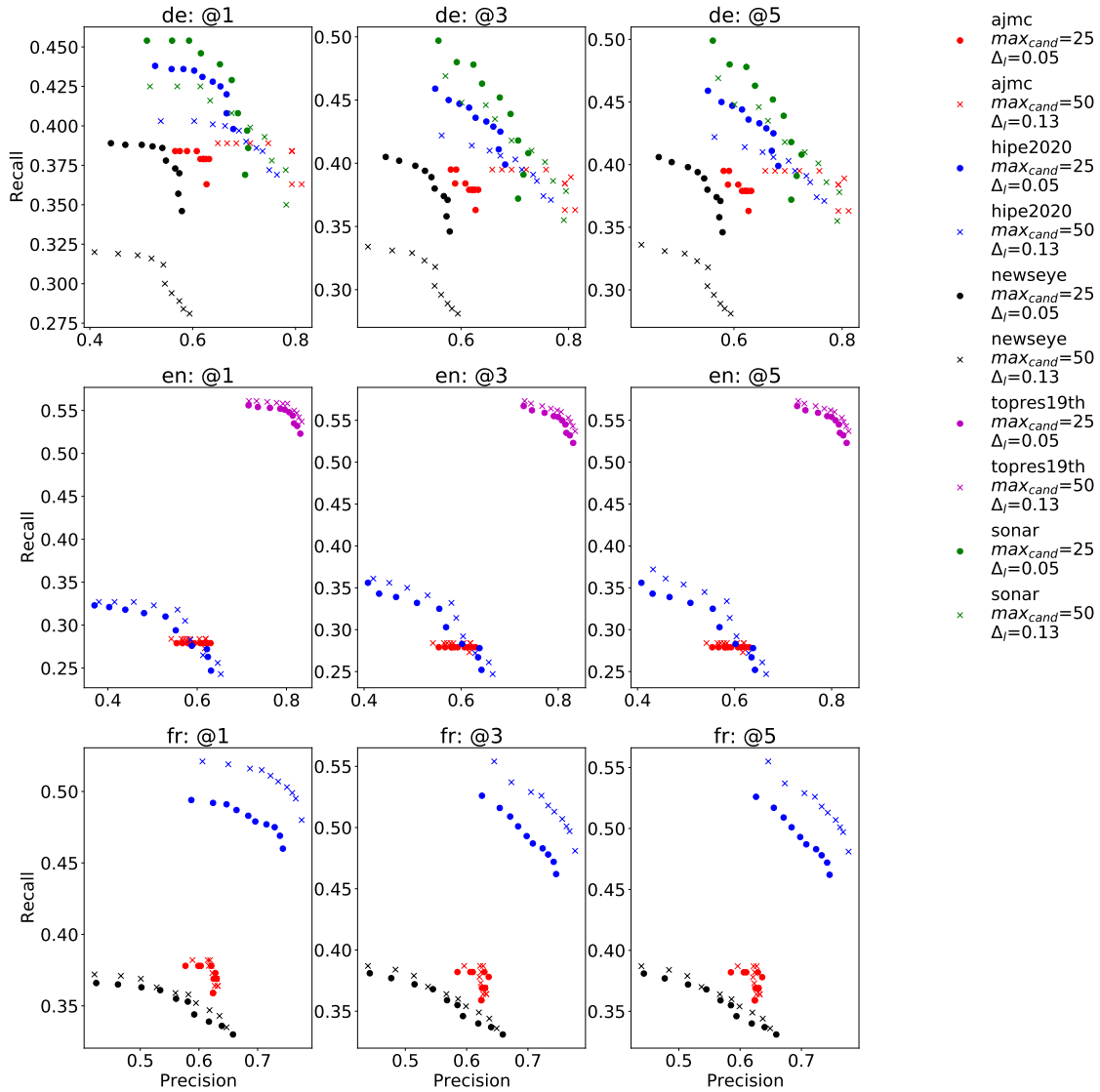


Figure 3: Influence of max_{cand} , Δ_l , and Δ_r on precision and recall on the test sets in the challenge. Color encodes the dataset whereas the marker shape encodes ($max_{cand} = 25, \Delta_l = 0.05$) versus ($max_{cand} = 50, \Delta_l = 0.13$).

much better coverage of relevant Wikidata-IDs in the test data. However there is still some variation in the coverage, and this variation cannot explain the performance differences between datasets.

Furthermore, the results for different datasets and languages show non-trivial qualitative impact of different parametrizations of our EL system. These findings need further investigation,

since a simple adaptation of the systems parameters does not offer a consistent improvement over all datasets and languages, and the systems response to a change in the hyperparameters is not identical for all the datasets.

In summary, the second edition of the CLEF HIPE Shared Task has again provided us with many useful insights into our system performance on diverse datasets, and we look forward to investigate the potential for further performance improvements according to these findings. In future work we aim to integrate visual embeddings into a multi-modal system (cf. [12]), and we will also further exploit prior known constraints (such as e.g. date of publication or birth/death dates of authors) for the purpose of EL.

References

- [1] K. Labusch, C. Neudecker, Named Entity Disambiguation and Linking on Historic Newspaper OCR with BERT., in: Conference and Labs of the Evaluation Forum (CLEF 2020), volume 2696, CEUR-WS Working Notes, 2020.
- [2] M. Ehrmann, M. Romanello, A. Flückiger, S. Clematide, Overview of CLEF HIPE 2020: Named Entity Recognition and Linking on Historical Newspapers, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2020, pp. 288–310.
- [3] M. Ehrmann, M. Romanello, A. Doucet, S. Clematide, Introducing the HIPE 2022 Shared Task: Named Entity Recognition and Linking in Multilingual Historical Documents, in: Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part II, Springer-Verlag, Berlin, Heidelberg, 2022, p. 347–354. URL: https://doi.org/10.1007/978-3-030-99739-7_44. doi:10.1007/978-3-030-99739-7_44.
- [4] T. M. Breuel, A. Ul-Hasan, M. A. Al-Azawi, F. Shafait, High-performance OCR for printed English and Fraktur using LSTM networks, in: 2013 12th international conference on document analysis and recognition, IEEE, 2013, pp. 683–687.
- [5] C. Wick, C. Reul, F. Puppe, Calamari - A High-Performance Tensorflow-based Deep Learning Package for Optical Character Recognition, Digital Humanities Quarterly 14 (2020).
- [6] G. Rehm, P. Bourgonje, S. Hegele, F. Kintzel, J. M. Schneider, M. Ostendorff, K. Zaczynska, A. Berger, S. Grill, S. Räuchle, J. Rauenbusch, L. Rutenburg, A. Schmidt, M. Wild, H. Hoffmann, J. Fink, S. Schulz, J. Seva, J. Quantz, J. Böttger, J. Matthey, R. Fricke, J. Thomsen, A. Paschke, J. A. Qundus, T. Hoppe, N. Karam, F. Weichhardt, C. Fillies, C. Neudecker, M. Gerber, K. Labusch, V. Rezanezhad, R. Schaefer, D. Zellhöfer, D. Siewert, P. Bunk, L. Pintscher, E. Aleynikova, F. Heine, QURATOR: Innovative Technologies for Content and Data Curation, CoRR abs/2004.12195 (2020). URL: <https://arxiv.org/abs/2004.12195>. arXiv:2004.12195.
- [7] K. Labusch, C. Neudecker, D. Zellhöfer, BERT for Named Entity Recognition in Contemporary and Historic German, in: Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019): Long Papers, German Society for Computational Linguistics

- & Language Technology, Erlangen, Germany, 2019, p. 1–9. URL: https://corpora.linguistik.uni-erlangen.de/data/konvens/proceedings/papers/KONVENS2019_paper_4.pdf.
- [8] S. Menzel, H. Schnaitter, J. Zinck, V. Petras, C. Neudecker, K. Labusch, E. Leitner, G. Rehm, Named Entity Linking mit Wikidata und GND—Das Potenzial handkuratierter und strukturierter Datenquellen für die semantische Anreicherung von Volltexten, *Qualität in der Inhaltserschließung* 70 (2021) 229–257.
- [9] E. Boros, E. L. Pontes, L. A. Cabrera-Diego, A. Hamdi, J. G. Moreno, N. Sidère, A. Doucet, Robust named entity recognition and linking on historical multilingual documents, in: *Conference and Labs of the Evaluation Forum (CLEF 2020)*, volume 2696, CEUR-WS Working Notes, 2020, pp. 1–17.
- [10] M. Ehrmann, M. Romanello, S. Najem-Meyer, A. Doucet, S. Clematide, Overview of HIPE-2022: Named Entity Recognition and Linking in Multilingual Historical Documents, in: A. Barrón-Cedeño, G. Da San Martino, M. Degli Esposti, F. Sebastiani, C. Macdonald, G. Pasi, A. Hanbury, M. Potthast, G. Faggioli, N. Ferro (Eds.), *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Thirteenth International Conference of the CLEF Association (CLEF 2022)*, Lecture Notes in Computer Science (LNCS), Springer, 2022.
- [11] Ehrmann, Maud and Romanello, Matteo and Flückiger, Alex and Clematide, Simon, CLEF-HIPE-2020 Shared Task Named Entity Datasets, 2020. [10.5281/zenodo.6046853](https://doi.org/10.5281/zenodo.6046853).
- [12] X. Wang, J. Tian, M. Gui, Z. Li, R. Wang, M. Yan, L. Chen, Y. Xiao, WikiDiverse: A Multimodal Entity Linking Dataset with Diversified Contextual Topics and Entity Types, arXiv preprint [arXiv:2204.06347](https://arxiv.org/abs/2204.06347) (2022).