Named Entity Recognition and Linking on Historical Newspapers: UvA.ILPS & REL at CLEF HIPE 2020

Vera Provatorova[1], Svitlana Vakulenko[1], Evangelos Kanoulas[1], Koen Dercksen[2], and Johannes M van Hulst[2]

[1] University of Amsterdam, Amsterdam, The Netherlands
[2] Radboud University, Nijmegen, The Netherlands

Abstract. This paper describes our submission to the CLEF HIPE 2020 shared task on identifying named entities in multi-lingual historical newspapers in French, German and English. The subtasks we addressed in our submission include coarse-grained named entity recognition, entity mention detection and entity linking. For the task of named entity recognition we used an ensemble of fine-tuned BERT models; entity linking was approached by three different methods: (1) a simple method relying on ElasticSearch retrieval scores, (2) an approach based on contextualised text embeddings, and (3) REL, a modular entity linking system based on several state-of-the-art components.

Keywords: Named Entity Linking · Named Entity Recognition

1 Introduction

Named entity identification is an important task in information extraction. Detecting, classifying and linking named entities helps to enable semantic search, which can be used for different domain applications, such as digital humanities [13]. One example is information retrieval from historical corpora. Identifying entities in historical documents poses several important challenges due to the nature of historical texts. These challenges include OCR errors in document scans, historical spelling variations and semantic shifts [12, 5]. This paper describes the submissions prepared by our joint team from the University of Amsterdam and Radboud University for the CLEF HIPE shared task. The main focus of CLEF HIPE is on systematic evaluation of named entity recognition and linking methods on multilingual diachronic historical data [6]. The shared task consists of several subtasks grouped into five bundles. Every team was allowed to submit one bundle per language, with the exception of bundle 5 (named entity linking given canonical mention spans), which was evaluated separately and could be combined with any other bundle.

Our submission targeted three of the subtasks in HIPE: (1) coarse-grained named entity recognition (NERC), (2) end-to-end named entity linking (NEL) using a modified NERC task for entity mention detection, and (3) named entity linking using mention spans provided by the organisers (NEL-only). Entity mention detection in this case was a supplementary task: it was not evaluated directly within the system submissions, but served as a preparation step for NEL in the setting of bundle 2, where entity mention boundaries were not given in the test data. In all the subtasks, we only considered the literal sense of the entities.

For the first phase of the shared task, we designed solutions for English, German and French languages within bundle 2, which included identifying, classifying and linking coarse-grained entities. For the second phase, bundle 5, we focused on one language only (English) and compared our results to the out-of-the-box tool, Radboud Entity Linker (REL) [10], as a competitive baseline.

2 Bundle 2: Named Entity Recognition and Linking

2.1 Experimental setup

Datasets and resources. The dataset provided by the CLEF HIPE organisers consists of diachronically organised digitised historical newspaper articles in English, German and French. The data is annotated using the standard inside–outside–beginning (IOB) format and presented as tab-separated values, where each row corresponds to a single token.

While validation datasets are provided for all of the three languages, training data are only available for German and French. To provide the token classification model with a sufficient amount of training data for English, we used CoNLL-03 [14] as an auxiliary dataset.

Approach. We consider both NERC and entity mention detection tasks as instances of the sequence classification task. For the NERC task, 5 entity types (org, pers, prod, loc, and time) form 11 classes when annotated in the IOB format: each of the types has its "B-" and "I-" labels corresponding to the tokens at the beginning and inside of an entity (e.g., "B-pers" and "I-pers"), while the "O" label marks the remaining tokens which are outside of named entities. For mention detection, 3 classes are considered: "B-entity", "I-entity", and "O". To perform sequence classification, we fine-tuned two pretrained BERT models [3] provided by the Hugging Face Transformers library [15]: `bert-base-cased` for English and `bert-base-multilingual-cased` for French and German. To improve robustness of the approach, we used a majority vote ensemble of 5 model instances per language fine-tuned on the training data with different numbers of epochs, as well as different random seed values, where $5 \leq num\_epochs \leq 9$ and $random\_seed = 42 + num\_epochs$.

To perform entity linking, we used ElasticSearch [4] to index all Wikidata entity labels and search for each of the entity mentions extracted from the input data to retrieve candidate entities. All the retrieved entities were included as candidates, without filtering on type. Candidate entity ranking was performed

based on ElasticSearch retrieval scores combined with several heuristics, preferring precise matching and shorter entity IDs (assuming that the entities with shorter IDs that were added to Wikidata earlier are typically more general and therefore more likely to be correct in many cases). We used the latest Wikidata dump from 9th of March 2020 which contains more than 55M entities. An important limitation of our approach is that it relied solely on the English-language labels, which is likely to hinder its performance on some of the named entities that vary across languages, such as "Geneva" in English versus "Genf" in German.

2.2 Results and discussion

The submissions were evaluated with the HIPE scorer, which is provided by the shared task organisers and available on github.[3] The scores achieved by our submissions on the NERC task are presented in Table 1.

Table 1. NERC-coarse results (literal sense, micro average)

	English						French						German					
	strict			fuzzy			strict			fuzzy			strict			fuzzy		
	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R
best_HIPE	.632	.623	.641	.786	.775	.797	.840	.831	.849	.921	.912	.931	.797	.790	.805	.878	.870	.886
UvA.ILPS	.473	.443	.508	.678	.635	.728	.686	.656	.719	.830	.794	.869	.526	.499	.556	.726	.689	.768
baseline_HIPE	.405	.531	.327	.562	.736	.454	.646	.693	.606	.769	.825	.721	.476	.643	.378	.585	.790	.464

The baseline provided by the HIPE organisers for the NERC-coarse task uses a traditional CRF sequence classification method. The top solution for all languages is developed by the L3i team, with extra layers added on top of several pre-trained BERT models and trained in a multi-task learning setting to minimize the impact of OCR-generated noise, historical spelling variations and other challenges specific to the data [2]. Our approach outperforms the baseline but achieves significantly lower results in comparison with the top solution. It shows that, while transformer-based approaches are a promising direction for named entity recognition, using a majority vote ensemble of fine-tuned models without any extra modifications is not likely to be sufficient for the setting of noisy historical data.

Table 2. End-to-end NEL results (literal sense, micro average)

	English						French						German					
	strict @1			relaxed @1			strict @1			relaxed @1			strict @1			relaxed @1		
	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R
best_HIPE	.531	.523	.539	.531	.523	.539	.594	.602	.598	.613	.622	.617	.534	.531	.538	.557	.553	.561
UvA.ILPS	.300	.249	.375	.300	.249	.375	.251	.352	.195	.252	.353	.196	.254	.241	.269	.264	.250	.279
baseline_HIPE	.220	.263	.239	.220	.263	.239	.206	.342	.257	.257	.358	.270	.173	.187	.180	.188	.203	.195

[3] https://github.com/impresso/CLEF-HIPE-2020-scorer

For the end-to-end NEL task, the HIPE baseline is AIDA-light trained on English Wikipedia. The best solution was submitted by the L3i team using entity embeddings trained on Wikipedia and Wikidata, combined with probabilistic mapping. The results achieved by our submissions are presented in Table 2 and compared with these two approaches.

For English and German, our submission scores above the baseline but far below the top solution, which is not surprising given the simplicity of our approach. For French the recall values of our submission are below the baseline. We assume that the main reason for this performance drop is due to the fact that most of the French entities could not be found in the English-only Wikidata index used in our system. We conclude that the bottleneck of our approach is the entity retrieval rather than entity mention detection.

3 Bundle 5: Named Entity Linking with Correct Mention Spans

3.1 Experimental setup

Datasets and resources. Our system runs were prepared using the same HIPE corpora as in bundle 2, with no extra training data. The algorithm designed for the first two runs used pre-trained contextualised Flair string embeddings [1] provided by the task organisers.

Methods. For the first two runs, candidate entity retrieval was done the same way as in bundle 2. To perform candidate entity ranking, we calculated cosine similarity between contextual embeddings of a sentence containing the target entity mention and a modified sentence, where the target entity mention is replaced with candidate entity description extracted from Wikidata. For example, if the target sentence is *"We went to London for a weekend"* and a candidate entity is *Q84* with the label *London* and the description *"capital and largest city of the United Kingdom"*, then the modified sentence would be *"We went to* ***capital and largest city of the United Kingdom*** *for a weekend"*.

The idea behind our approach resides upon two basic assumptions: (1) Wikidata entity descriptions are semantically similar to the corresponding entity labels, and (2) contextualised string embeddings capture similarity between entity descriptions and entity labels. After calculating the cosine similarity score, it is multiplied by the Levenshtein similarity ratio between target and candidate entity labels to prefer precise matching where possible. In the example above, if one of the candidates is *Q23306: Greater London* then its score would be multiplied by *sim('London', 'Greater London') = 0.6*, while the score for *Q84: London* would remain the same, as *sim('London', 'London') = 1*. The similarity ratio was calculated using the FuzzyWuzzy string matching library [7].

After using the resulting score to rank the list of candidate entities, a NIL value is inserted to the list before the first candidate that has a score below threshold. We chose the threshold value of 0.7 after tuning this parameter on the

development set. For submission 2 only, we added historical spelling variations to the step of candidate retrieval using Natas library that performs historical normalisation via neural machine translation [9].

The third run was prepared using REL [10] – a modular system that is based on several state-of-the-art components, available as a Python library as well as a web API[4]. Entity linking in REL is divided into three components: (i) mention detection, (ii) candidate selection, and (iii) entity disambiguation. For this submission, mention detection was skipped since the mention spans were already provided by the organisers as the ground truth. Candidate selection consists of retrieving seven candidates for each mention. The first four candidates are retrieved based on the co-occurence probability of entities given a specific mention (a so called $p(e|m)$ *index*). The remaining three are selected based on their contextual similarity to the mention in an embedding space.

Entity disambiguation decisions are made by combining local compatibility (which includes prior importance and contextual similarity) and coherence with the other entity linking decisions in a document (global context).

3.2 Results and discussion

Run 1: Baseline. While the results @1 are below the HIPE baseline (Table 3), the performance @3 and @5 is better (Table 4). Similar results were achieved on the development set: while the correct entity would often make it to the top-5 or top-3 of the ranked candidate list, it was rarely selected by the algorithm as the most relevant answer, and the difference between candidate scores was usually small. The algorithm was not directly optimised for top-1 candidate selection. Another obstacle for the algorithm was NIL detection: as 30% of the mentions were not linkable [6], simply adding the NIL value to the ranked list of candidates based on the fixed threshold value was not a sufficient approach and resulted in an overwhelming number of false positives.

Table 3. Named entity linking results (English, literal sense)

	strict @1			fuzzy @1		
	F	P	R	F	P	R
best_HIPE	.633	.685	.658	.633	.685	.658
Run #3 REL	.593	.607	.580	.593	.607	.580
Run #1 Baseline	.367	.365	.369	.367	.365	.369
Run #2 Historical	.348	.344	.353	.348	.344	.353
baseline_HIPE	.506	.506	.506	.506	.506	.506

[4] https://github.com/informagi/REL

Table 4. NEL-only results @3 and @5 (English, literal sense)

	@3			@5		
	F	P	R	F	P	R
Run #1 Baseline	.463	.467	.465	.552	.557	.555
Run #2 Historical	.451	.463	.457	.540	.555	.548

Run 2: Historical normalisation. Adding extra candidate entities by means of historical normalisation in the second submission has resulted in more false positives and slightly decreased overall performance in comparison to the first submission. A likely explanation is that the normalisation algorithm was focusing on infrequent historical spellings [9], most of which are not likely to be present in the HIPE dataset.

Run 3: REL. REL performs very well and takes the second place in the scoring table, which is rather remarkable for an out-of-the box linking system. We showed that REL provides a strong baseline for the NEL task on historical documents, demonstrating the state-of-the-art performance that can be reached without accounting for additional properties, such as OCR errors and language change.

4 Conclusion and future work

Our contributions within the CLEF HIPE shared task approached coarse-grained named entity recognition (NERC) and two settings of entity linking: end-to-end and NEL-only. The results for NERC show that although fine-tuning BERT models for sequence classification is enough to outperform the baselines for all three languages, achieving top performance requires extra modifications in order to deal with the challenges specific to historical data. The NEL results show that, while using an embedding-based approach that takes historical spelling variations into account is better than relying solely on ElasticSearch retrieval scores, this approach is clearly outperformed by REL, as well as by many other solutions – mostly due to its poor performance on NIL prediction and an overwhelming number of false positives on the candidate selection step. REL, in its turn, proves very efficient in the setting of the shared task, even without specifically addressing the challenges of the historical data.

There are several possible directions for future work considering all the subtasks that we approached in the context of the shared task:

Entity recognition and classification. Some examples of the ways to achieve improvements over the state-of-the-art sequence classification methods within the given task setup include (i) performing a more extensive parameter search for the Transformer models; (ii) fine-tuning more advanced pre-trained models (such as RoBERTa [11]), and (iii) reducing the impact of the noise in the training data by using OCR correction algorithms, such as [8].

Entity linking. Since the task of entity linking consists of several steps, including candidate generation and entity disambiguation, we see further opportunities for improvement on each of these steps. Firstly, candidate generation can be improved to increase recall. One of the ways to achieve this goal is to use OCR correction as a pre-processing step in the algorithm. Secondly, entity disambiguation should be improved upon in order to increase precision by decreasing the number of false positives. We consider graph-based disambiguation methods as a promising research direction. Thirdly, using entity types as features instead of only relying on mention boundaries could also improve entity disambiguation in the end-to-end setting.

Acknowledgements

This research was supported by the NWO Innovational Research Incentives Scheme Vidi (016.Vidi.189.039), the NWO Smart Culture - Big Data / Digital Humanities (314-99-301), the H2020-EU.3.4. - SOCIETAL CHALLENGES - Smart, Green And Integrated Transport (814961) the Google Faculty Research Awards program. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

References

1. Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., Vollgraf, R.: Flair: An easy-to-use framework for state-of-the-art nlp. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations). pp. 54–59 (2019)
2. Boros, E., Linhares Pontes, E., Cabrera-Diego, L.A., Hamdi, A., Moreno, J.G., Sidère, N., Doucet, A.: Robust Named Entity Recognition and Linking on Historical Multilingual Documents. In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) CLEF 2020 Working Notes. Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum. CEUR-WS (2020)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
4. Divya, M.S., Goyal, S.K.: Elasticsearch: An advanced and quick search technique to handle voluminous data. Compusoft **2**(6), 171 (2013)
5. Ehrmann, M., Colavizza, G., Rochat, Y., Kaplan, F.: Diachronic Evaluation of NER Systems on Old Newspapers. In: Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)). pp. 97–107. Bochumer Linguistische Arbeitsberichte (2016), https://infoscience.epfl.ch/record/221391?ln=en
6. Ehrmann, M., Romanello, M., Flückiger, A., Clematide, S.: Overview of CLEF HIPE 2020: Named Entity Recognition and Linking on Historical Newspapers. In: Arampatzis, A., Kanoulas, E., Tsikrika, T., Vrochidis, S., Joho, H., Lioma, C., Eickhoff, C., Névéol, A., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 11th International Conference of the CLEF Association (CLEF 2020). Lecture Notes in Computer Science (LNCS), vol. 12260. Springer (2020)

7. Gonzalez, J., Rodrigues, P., Cohen, A.: Fuzzywuzzy: Fuzzy string matching in python (2017)
8. Hämäläinen, M., Hengchen, S.: From the paft to the fiiture: a fully automatic NMT and word embeddings method for OCR post-correction. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019). pp. 431–436. INCOMA Ltd., Varna, Bulgaria (Sep 2019), https://www.aclweb.org/anthology/R19-1051
9. Hämäläinen, M., Säily, T., Rueter, J., Tiedemann, J., Mäkelä, E.: Revisiting NMT for normalization of early English letters. In: Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature. pp. 71–75. Association for Computational Linguistics, Minneapolis, USA (Jun 2019), https://www.aclweb.org/anthology/W19-2509
10. van Hulst, J.M., Hasibi, F., Dercksen, K., Balog, K., de Vries, A.P.: REL: An entity linker standing on the shoulders of giants. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '20, ACM (2020)
11. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
12. Piotrowski, M.: Natural language processing for historical texts. Synthesis lectures on human language technologies **5**(2), 1–157 (2012)
13. Provatorova, V., Kanoulas, E., Carlgren, A., Dupré, S., Hendriksen, M.: Art DATIS: Improving search in multilingual corpora to support art historians. Digital Humanities Benelux '19 (2019)
14. 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. pp. 142–147 (2003), https://www.aclweb.org/anthology/W03-0419
15. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al.: Huggingface's transformers: State-of-the-art natural language processing. ArXiv pp. arXiv–1910 (2019)