Entities, Dates, and Languages: Zero-Shot on Historical Texts with T0

Anonymous ACL submission

Abstract

In this work, we explore whether the recently demonstrated zero-shot abilities of the T0 model extend to Named Entity Recognition for out-of-distribution languages and time periods. Using a historical newspaper corpus in 3 languages as test-bed, we use prompts to extract possible named entities. Our results show that a naive approach for prompt-based zero-shot multilingual Named Entity Recognition is error-prone, but highlights the potential of such an approach for historical languages lacking labeled datasets. Moreover, we also find that T0-like models can be probed to predict the publication date and language of a document, which could be very relevant for the study of historical texts.

1 Introduction

This paper lies at the focal point of three orthogonal advances. First, the recent surge in GLAM[1]-led digitisation efforts (Terras, 2011), open citizen science (Haklay et al., 2021) and the expansive commodification of data (Hey and Trefethen, 2003), have enabled a new mode of historical inquiry that capitalises on the 'big data of the past' (Kaplan and Di Lenardo, 2017). Second, the 2017 breakthrough that was the transformer architecture (Vaswani et al., 2017) has led to the so-called ImageNet moment of Natural Language Processing (Ruder, 2018) and brought about unprecedented progress in transfer-learning (Raffel et al., 2020), few-shot learning (Schick and Schütze, 2021), zero-shot learning (Sanh et al., 2021), and prompt-based learning (Le Scao and Rush, 2021) for natural language. Third, the growing popularity of prompt-based methods (Liu et al., 2021) has resulted in a new paradigm for training and fine-tuning Large Language Models (LLM) as well as novel applications in Named Entity Recognition (NER) (Liu et al., 2022).

[1]Galleries, libraries, archives, and museums.

NER for historical texts has been the focus of a growing body of research, most recently surveyed by Ehrmann et al. (2021). Both NER and the related task of Entity Linking can enhance our ability to search and navigate digitised historical materials (Neudecker et al., 2014; Kim and Cassidy, 2015). However, applying NER to historical texts poses a number of challenges, including those due to errors in Optical Character Recognition (OCR) (Ehrmann et al., 2021; Hamdi et al., 2019; Boros et al., 2020) and domain transfer (Baptiste et al., 2021). To advance research in this area, an increasing number of datasets have been created to support the development and evaluation of NER approaches in historical text (Neudecker, 2016; Ehrmann et al., 2020, 2022)

In this paper, we examine the zero-shot abilities of T0— a prompt-based LLM developed as part of the BigScience project for open research (Sanh et al., 2021)— on the challenging task of historical NER[2]. This endeavour had two main hurdles: (1) the model was neither trained to recognize entities, nor was it ever tested on that task; (2) our evaluation dataset was out-of-distribution, containing both multilingual and historical data. To better contextualize the results of our experiments, we also run zero-shot prompt-based probing (Zhong et al., 2021) to assess T0's broader ability of extracting factual knowledge about two key factors in our experiment, that is, language variation and historical variation in the dataset.

2 Experimental setup

2.1 Data description

Our data comes from version 1.4 of the CLEF-HIPE[3] 2020 open-access dataset[4]: an OCR'ed newspaper corpus annotated for NER (Ehrmann

[2]Code and data will be released alongside the paper.

[3]Conference and Labs of the Evaluation Forum - Identifying Historical People, Places and other Entities.

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

	English			German			French		
Time period	#Documents	#Tokens	NE%	#Documents	#Tokens	NE%	#Documents	#Tokens	NE%
1790-1810	10	4143	3.1	13	6735	4.6	14	8550	4.4
1810-1830	15	4697	3.4	13	8049	2.6	10	12 440	5.0
1830-1850	9	3974	4.0	19	15 601	2.8	10	11 659	3.9
1850-1870	0	0	-	21	16 021	3.8	9	10 321	3.9
1870-1890	7	2202	1.9	16	17 181	3.7	15	16 272	4.2
1890-1910	12	4509	2.9	12	12 829	4.3	19	16 874	4.6
1910-1930	13	5499	3.1	13	18 134	3.3	30	30 403	3.8
1930-1950	3	520	4.2	29	24 566	5.7	32	35 962	4.2
Total	69	25 544	3.2	136	119 116	4.0	139	142 481	4.2

Table 1: Data description: splits by date and language of the CLEF-HIPE 2020 dataset.

et al., 2020). It contains Swiss and Luxembourgish newspapers from 1790 to 2010, in English, German and French. For our experiment, we use only entities of coarse type, according to their literal sense. Coarse entity types in the CLEF-HIPE 2020 dataset are persons (PERS), locations (LOC), organizations (ORG), dates (TIME) and products (PROD, which includes media and doctrines).

We use it in its entirety as test set[5], split by language and date (using 20 years time intervals[6] see Table 1). Each language dataset is relatively balanced between 1810 and 1910, with English containing between 2,202 and 4,697 tokens per split with the exception of one split (1850-1870 English) for which there are no tokens. German contains between 6,735 and 12,829 tokens, and French contains between 8,550 and 16,874 tokens. The end periods contain on average more tokens for German and French. Overall, the dataset contains 3.8% of named entities (from 1.9 to 5.6%, depending on time periods and datasets). The most balanced dataset across time periods is the French one (between 3.8 and 4.6% NE).

2.2 Model description

In our experiments, we use the T0++ variant of the T0 language model (Sanh et al., 2021), based on the LM-adapted T5 model (Lester et al., 2021), itself a variant of the T5 model (Raffel et al., 2020), which further pretrains the original encoder-decoder architecture of T5 with an autoregressive language modeling objective.[7] Crucially, this pretraining is done using a prompt-based training setup, in which training examples are transformed into prompts using a variety of crowd-sourced prompt templates. This setup allows T0 to perform few-shot and zero-shot learning when presented with new prompts for an previously unseen task.

2.3 Experimental setup

Our goal in this paper is to see if and how state-of-the-art language models can be used for historical NLP tasks, with minimal modifications and fine-tuning.[8] As such, we choose to use a 'naive' approach, by directly asking the model which named entities a given sentence contains. To do so, we first design prompts for each named entity type (e.g. see Table 2). For each sentence in the dataset, we then 1) use all the generation prompts to determine if the sentence contains named entities of each entity type [9]; 2) filter the model's answer to keep only tokens that are actually in the input sentence, keeping the entity covering the longer span in case of nested entities; and 3) ask a disambiguation question if needed (if a token was assigned to multiple entities by the model). Results are stored at each step.

We then evaluate the results and conduct two additional experiments to better understand the impact of the dataset language and time period on the performance of the LM.

3 Results

3.1 Limitations

Results reveal limitations in our proposed approach. First, T0 exhibits a clear tendency to produce non-

[5]We mix the original training and validation sets to constitute our test set.

[6]We chose 20-year spans as the smallest time range producing somewhat balanced splits.

[7]The added specific pretraining of T0 uses a set of 11 varied tasks represented by a total of 55 datasets.

[8]Ecological concerns and funding inequalities raise considerations on how to best use already existing models for lower-resourced tasks, and with spending as little further computing power in fine-tuning as possible (Bender et al., 2021).

[9]For PROD entities, the generation prompt explicitly mentioned *media* and *doctrines*, as we regarded the word *product* as too generic to return an accurate answer from T0.

Entity	Step (1) Generation prompt
PERS	Input: **<sentence>\n** In input, what are the names of person? Separate answers with commas.
LOC	Input: **<sentence>\n** In input, what are the names of location? Separate answers with commas.
PROD	Input: **<sentence>\n** In input, what are the names of media or doctrine? Separate answers with commas.
Entities	**Step (3) Disambiguation prompt**
PERS, LOC	Input: **<sentence>\n** In input, is **<entity>** a person or a location? Give only one answer.
Fact	**Factual probing prompts**
Language	**<sentence>\n** Q:Name the language of the previous sentence.**\n**A:
Date	In which year is the following text likely to have been published: text: **<text>**

Table 2: Example prompts for generation and disambiguation (Sec. 2.3), as well as factual probing (Sec. 4).

empty outputs regardless of the presence or absence of named entities in the input: none of the prompts generates an empty answer. This is especially visible for the entity PROD, for which T0 answers over 55% of the queries with the name of the entity itself (e.g. either *media* or *doctrine*) rather than with any other token from the input sentence. Second, adequately matching T0's output with tokens in the input sentence proved difficult. Even when T0 generates an answer semantically very close to the correct token in the sentence, differences in spelling prevent the algorithm from correctly associating T0's answer with said token in the input sentence. This problem is inherent to the nature of our dataset: frequent OCR errors generate unpredictable variations in 'gold' word spelling (including spacing between words and letters or diacritics variation), which are automatically corrected by T0 during its predictions,[10] which negatively affects our ability to automatically match its answers with corresponding tokens in the sentence. In other instances, the model translated words from French and German into English. Further experiments might need to mitigate language variety by adding input text to the prompt, to help the model correctly assess the language in which it must answer. As all answers predicted are considered strictly incorrect, the algorithm never enters its disambiguation phase. We therefore analyse non disambiguated results.

3.2 Evaluation

To evaluate proximity between predictions and gold, we compare 'gold' tokens with predicted tokens using normalized Levenshtein distance (Levenshtein, 1966),[11] using this metric as a proxy to identify best predictions for each entity query in each sentence. For a given example, we define (1) the true positive as the prediction with the shortest Levenshtein distance from the gold; (2) false positives as predictions of entities that are not actually present in the input sentence; and (3) false negatives as predictions that have longer Levenshtein distance to the gold tokens (i.e. predictions that would have failed to identify entity tokens in the sentence). Precision and F1-score are relatively low, especially for PROD entities, which were the most difficult to define in terms of text prompts. Higher values for recall are due to the fact that increasing the Levenshtein threshold makes it more likely to find an acceptable answer among those generated by T0. Unsurprisingly, the highest increase is found in TIME entities (dates have fixed formats, which makes it more likely to find an acceptable distance between predictions and correct tokens). Precision scores for each entity type are shown in Figure 1 (see Fig. 3 in Appendix for recall and F1-score). The results of our experiment suggest that, although T0 struggles to return exact matches of the entities in the input sentence, it is still capable of generating answers that are semantically close to the correct tokens.

After manually inspecting the dataset and its numerous OCR artifacts, we choose 0.4 as a reasonable heuristic of close semantic similarity between T0's output and gold tokens. We find that using a threshold of 0.4 prevents the apparition of false positives, and therefore we use it to analyze differences between languages and between historical periods within the dataset. With respect to variations across languages, we observe that the precision of predictions in English does not have a clear hedge over precision in French and German (Fig. 2; see also Fig. 4 in Appendix). This is unexpected, as T0 should display considerable bias towards English,

[10] E.g. Respelling words that were garbled due to noisy OCR.

[11] Normalization was done with regard to the length of the longest token (predicted or correct), and results were kept below a threshold. We tried 0.0, 0.1, 0.2, 0.3, 0.4 and 0.5.

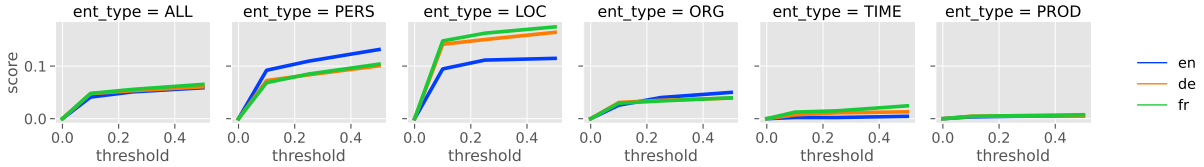

Figure 1: Precision at different Levenshtein distance thresholds.

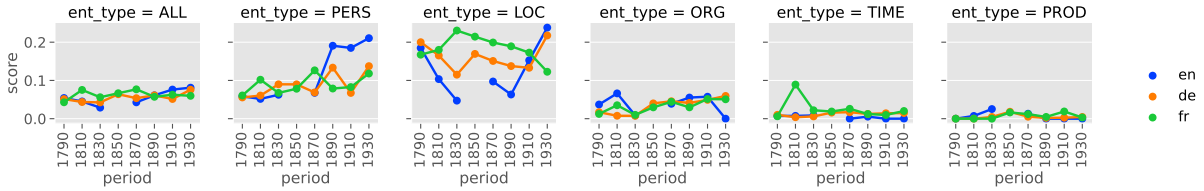

Figure 2: Precision at Levenshtein threshold 0.4 across periods.

which constitutes most of its training data. With respect to variations across periods, we observe an improvement in precision (and F1-score) for PERS and LOC entities in English texts from 1850s onwards (Fig. 3; for recall and F1-score, see Fig. 5 in Appendix), when for other entities and languages, precision and F1-score are either stable or show a downward trend (e.g. LOC in German)[12]. Variations in recall cannot be reduced to clear trends, but they are particularly erratic in English texts. A possible explanation could be that T0 is more sensitive to English text inputs, and therefore outputs a higher or lower number of irrelevant answers based on the specific content of each input sentence.

4 Prompt-based factual probing

In addition to our main experiment on NER, we run two further experiments to assess T0's ability to do inference in a multilingual setting and to identify historical variation in textual corpora.

Probing for language To gauge T0's ability to reason in a multilingual setting, we test the model's language identification ability. To that end, we use a trilingual[13] subset of the WiLI-2018 - Wikipedia Language Identification dataset (Thoma, 2018) and prompt the model on language (Table 2). We find that the model is able to correctly classify 83% of French sentences, 74.1% of German sentences, but only 35.4% of English sentences. The previously mentioned potential sensitivity of the model to its own mother tongue might explain this result.

[12]The absence of documents in the 1850-1870 English split explains the missing values for English in that period.

[13]French, German, and English; 1000 sentences each.

Probing for publication date To assess T0's treatment of historical text, we study how well it predicts the likely date of publication for a piece of text from our test dataset by prompting on publication date (Table 2).

	Absolute errors	
Languages	Mean	Median
English	40.48	30.0
German	40.11	32.0
French	55.25	48.0

Table 3: Date prediction results.

Table 3 shows the prediction errors. Subtle language change can occur in a measurable way in as short a period as a decade (Juola, 2003), and therefore a median absolute error of 30 suggests that T0 is good in predicting publication dates. We notice some variation in performance between different languages, with French performing slightly worse on both metrics (possibly because it belongs to a different language family from English, contrary to German).

5 Conclusion

We have presented our experiment to evaluate T0 for zero-shot historical NER, as well as on the prediction of language and publication date of historical texts. Our results show that historical texts present additional challenges for zero-shot NER (especially because historical datasets often include noisy OCR), but that T0 can however be used as is for language and date prediction. Next steps will be experimenting on different prompts and matching methods, as well as testing few-shot NER.

References

Blouin Baptiste, Benoit Favre, Jeremy Auguste, and Christian Henriot. 2021. Transferring Modern Named Entity Recognition to the Historical Domain: How to Take the Step? In *Workshop on Natural Language Processing for Digital Humanities (NLP4DH)*, Silchar (Online), India.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 610–623, New York, NY, USA. Association for Computing Machinery.

Emanuela Boros, Ahmed Hamdi, Elvys Linhares Pontes, Luis Adrián Cabrera-Diego, Jose 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*, pages 431–441, Online. Association for Computational Linguistics.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2021. Named entity recognition and classification on historical documents: A survey. *CoRR*, abs/2109.11406.

Maud Ehrmann, Matteo Romanello, Antoine Doucet, and Simon Clematide. 2022. HIPE-2022 shared task named entity datasets.

Maud Ehrmann, Matteo Romanello, Alex Flückiger, and Simon Clematide. 2020. Extended Overview of CLEF HIPE 2020: Named entity processing on historical newspapers. In *CLEF 2020 Working Notes. Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum*, volume 2696, page 38, Thessaloniki, Greece. CEUR-WS.

Fred Gibbs and Trevor Owens. 2013. The hermeneutics of data and historical writing. In Kristen Nawrotzki and Jack Dougherty, editors, *Writing History in the Digital Age*, pages 159–172. University of Michigan Press.

Frederick W Gibbs. 2016. New forms of history: Critiquing data and its representations. *The American Historian*, 7:31–36.

Muki Haklay, Dilek Fraisl, Bastian Tzovaras, Susanne Hecker, Margaret Gold, Gerid Hager, Luigi Ceccaroni, Barbara Kieslinger, Uta Wehn, Sasha Woods, Christian Nold, Bálint Balázs, Marzia Mazzonetto, Simone Ruefenacht, Lea Shanley, Katherin Wagenknecht, Alice Motion, Andrea Sforzi, Dorte Riemenschneider, and Katrin Vohland. 2021. Contours of citizen science: a vignette study. *Royal Society Open Science*, 8:202108.

Ahmed Hamdi, Axel Jean-Caurant, Nicolas Sidère, Mickaël Coustaty, and Antoine Doucet. 2019. An analysis of the performance of named entity recognition over ocred documents. *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 333–334.

Tony Hey and Anne Trefethen. 2003. *The Data Deluge: An e-Science Perspective*, chapter 36. John Wiley & Sons, Ltd.

Patrick Juola. 2003. The time course of language change. *Computers and the Humanities*, 37(1):77–96.

Frédéric Kaplan and Isabella Di Lenardo. 2017. Big data of the past. *Frontiers Digit. Humanit.*, 4:12.

Margaret MR Kellow. 1999. Erasing slavery: Memory, history, and race in new england. *Reviews in American History*, 27(4):526–533.

Sunghwan Mac Kim and Steve Cassidy. 2015. Finding names in Trove: Named entity recognition for Australian historical newspapers. In *Proceedings of the Australasian Language Technology Association Workshop 2015*, pages 57–65, Parramatta, Australia.

Teven Le Scao and Alexander Rush. 2021. How many data points is a prompt worth? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2627–2636, Online. Association for Computational Linguistics.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Vladimir Iosifovich Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Andy T. Liu, Wei Xiao, Henghui Zhu, Dejiao Zhang, Shang-Wen Li, and Andrew Arnold. 2022. QaNER: Prompting question answering models for few-shot named entity recognition.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing.

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'16)*, pages 4348–4352, Portorož, Slovenia. European Language Resources Association (ELRA).

Clemens Neudecker, Lotte Wilms, Willem Jan Faber, and Theo van Veen. 2014. Large-scale refinement of

digital historic newspapers with named entity recognition. In *IFLA Congress 2014 – Digital Transformation and the Changing Role of News Media in the 21st Century*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Christelle Ram. 2020. *Black historical erasure: A critical comparative analysis in Rosewood and Ocoee*. Ph.D. thesis, Rollins College.

Sebastian Ruder. 2018. NLP's ImageNet moment has arrived. https://ruder.io/nlp-imagenet/.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M. Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Stella Biderman, Leo Gao, Tali Bers, Thomas Wolf, and Alexander M. Rush. 2021. Multitask prompted training enables zero-shot task generalization. *CoRR*, abs/2110.08207.

Timo Schick and Hinrich Schütze. 2021. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

Michelle A Stanley. 2021. *Beyond erasure: Indigenous genocide denial and settler colonialism*. Routledge.

Melissa M. Terras. 2011. The rise of digitization. In Ruth Rikowski, editor, *Digitisation Perspectives*, pages 3–20. Sense Publishers, Rotterdam.

Martin Thoma. 2018. WiLI-2018 - Wikipedia Language Identification database.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Zexuan Zhong, Dan Friedman, and Danqi Chen. 2021. Factual probing is [MASK]: Learning vs. learning to recall. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5017–5033, Online. Association for Computational Linguistics.

A Broader Impacts Statement

In this paper, we take exploratory first steps toward instrumentalising the T0 large language model on the task of historical NER. We deem it appropriate to briefly discuss the ethical considerations that are implied by such a usage. First, if a model can be used in a context for which it was not explicitly intended for, it stands to reason that it can be misused in that same context: while recognizing entities in historical texts might at first glance seem innocuous, numerous studies focused on BIPOC representation in history have shown that this is not the case, as some marginalized groups tend to suffer from history erasure (Kellow, 1999; Ram, 2020; Stanley, 2021). Second, the automation and scaling of historical inquiry could potentially lead to unreflected (mis)interpretations of the past (Gibbs and Owens, 2013; Gibbs, 2016). Third, the experimental nature of prompt-based inference could lead to a considerable carbon footprint, owing to the trial-and-error nature of manual prompt calibration, though this cost would still be lower than training a new model from scratch or fine-tuning an existing LM (see footnote 8).

B Full scores of Levenshtein distance

The figures below and in the next page provide full results of evaluation on Levenshtein distance, including precision, recall and F1-score at different thresholds, at threshold 0.4, and across different time periods in the CLEF-HIPE 2020 dataset.

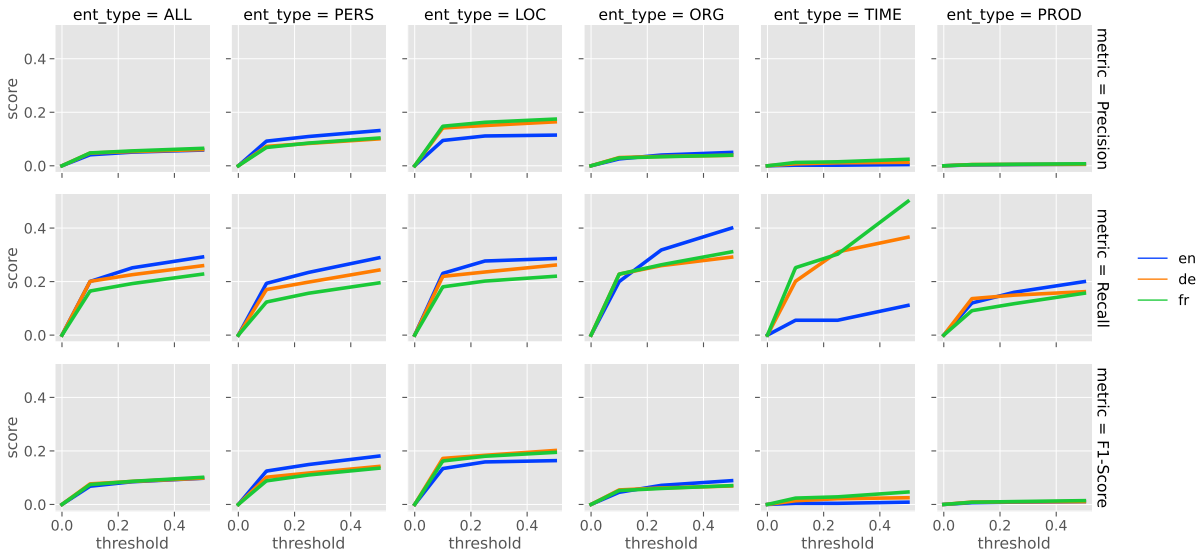

Figure 3: Precision, recall and F1-score at different Levenshtein distance thresholds.

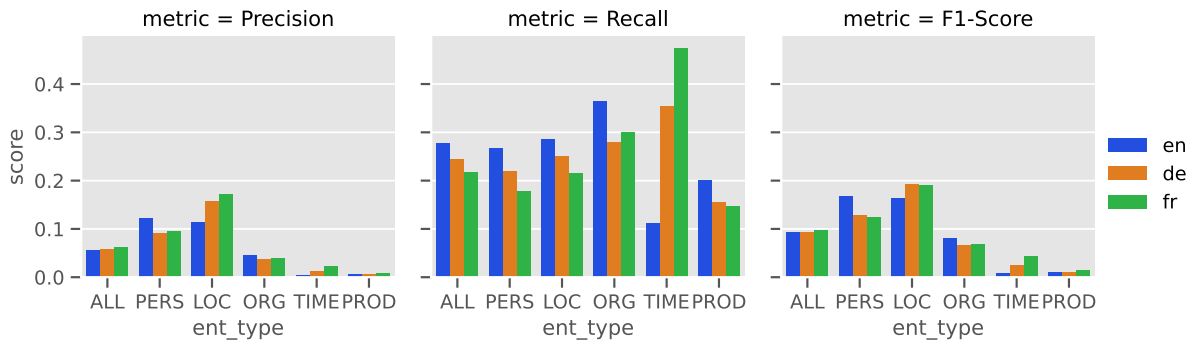

Figure 4: Precision, recall and F1-score by entity type at Levenshtein distance threshold 0.4.

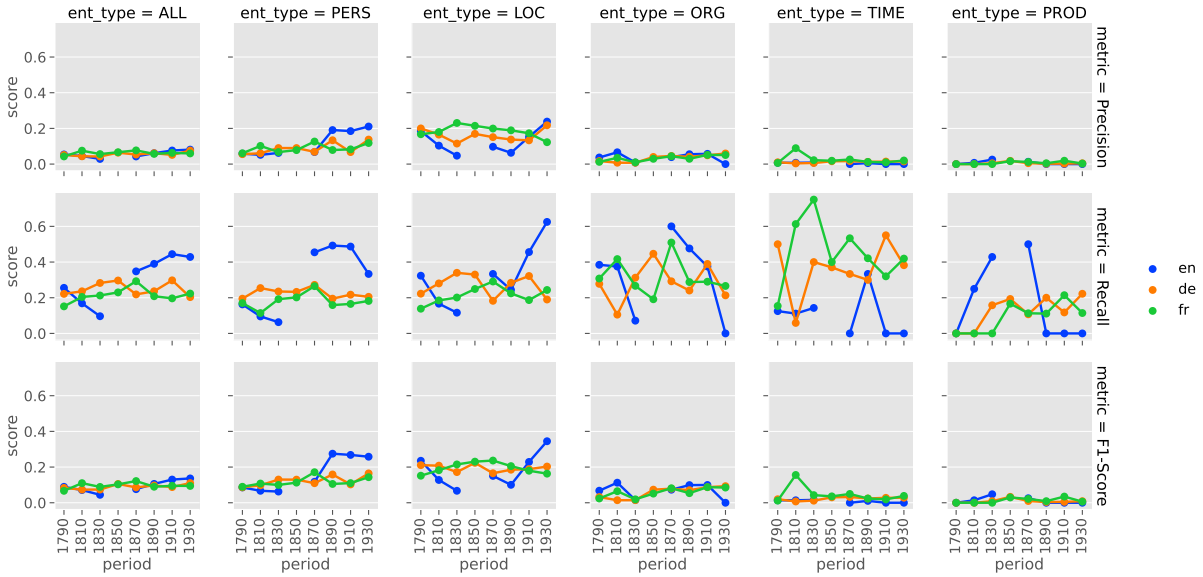

Figure 5: Precision, recall and F1-score at Levenshtein threshold 0.4 across periods.