








# Injecting Temporal-Aware Knowledge in Historical Named Entity Recognition

Carlos-Emiliano González-Gallardo<sup>1</sup>(✉) , Emanuela Boros<sup>1</sup> ,  
Edward Giamphy<sup>1,2</sup> , Ahmed Hamdi<sup>1</sup> , José G. Moreno<sup>1,3</sup> ,  
and Antoine Doucet<sup>1</sup> 

<sup>1</sup> University of La Rochelle, L3i, 17000 La Rochelle, France  
{carlos.gonzalez\_gallardo, emanuela.boros, ahmed.hamdi,  
antoine.doucet}@univ-lr.fr

<sup>2</sup> Preligens, 75009 Paris, France  
edward.giamphy@preligens.com

<sup>3</sup> University of Toulouse, IRIT, 31000 Toulouse, France  
jose.moreno@irit.fr

**Abstract.** In this paper, we address the detection of named entities in multilingual historical collections. We argue that, besides the multiple challenges that depend on the quality of digitization (e.g., misspellings and linguistic errors), historical documents can pose another challenge due to the fact that such collections are distributed over a long enough period of time to be affected by changes and evolution of natural language. Thus, we consider that detecting entities in historical collections is time-sensitive, and explore the inclusion of temporality in the named entity recognition (NER) task by exploiting temporal knowledge graphs. More precisely, we retrieve semantically-relevant additional contexts by exploring the time information provided by historical data collections and include them as mean-pooled representations in a Transformer-based NER model. We experiment with two recent multilingual historical collections in English, French, and German, consisting of historical newspapers (19C-20C) and classical commentaries (19C). The results are promising and show the effectiveness of injecting temporal-aware knowledge into the different datasets, languages, and diverse entity types.

**Keywords:** Named entity recognition · Temporal information extraction · Digital humanities

## 1 Introduction

Recent years have seen the delivery of an increasing amount of textual corpora for the Humanities and Social Sciences. Representative examples are offered by the digitization of the gigantic *Gallica* collection by the National Library of France<sup>1</sup> and the *Trove* online Australian library<sup>2</sup>, database aggregator and

<sup>1</sup> <https://gallica.bnf.fr/>.

<sup>2</sup> <https://trove.nla.gov.au/>.

service of full-text documents, digital images and data storage of digitized documents. Access to this massive data offers new perspectives to a growing number of disciplines, going from socio-political and cultural history to economic history, and linguistics to philology. Billions of images from historical documents including digitized manuscript documents, medieval registers and digitized old press are captured and their content is transcribed, manually through dedicated interfaces, or automatically using optical character recognition (OCR) or handwritten text recognition (HTR). The mass digitization process, initiated in the 1980s with small-scale internal projects, led to the “rise of digitization”, which grew to reach a certain maturity in the early 2000s with large-scale digitization campaigns across the industry [12, 16]. As this process of mass digitization continues, increasingly advanced techniques from the field of natural language processing (NLP) are dedicated to historical documents, offering new ways to access full-text semantically enriched archives [33], such as NER [4, 10, 19], entity linking (EL) [26] and event detection [5, 32].

However, for developing such techniques, historical collections present multiple challenges that depend either on the quality of digitization, the need to handle documents deteriorated by the effect of time, the poor quality printing materials or inaccurate scanning processes, which are common issues in historical documents [20]. Moreover, historical collections can pose another challenge due to the fact that documents are distributed over a long enough period of time to be affected by language change and evolution. This is especially true in the case of Western European languages, which only acquired their modern spelling standards roughly around the 18th or 19th centuries [29]. With existing collections [12, 15, 16] providing such metadata as the year of publication, we propose to take advantage of the temporal context of historical documents in order to increase the quality of their semantic enrichment. When this metadata is not available, due to the age of the documents, the year has often been estimated and a new NLP task recently emerged, aiming to predict a document’s year of publication [36].

NER corresponds to the identification of entities of interest in texts, generally of the type person, organization, and location. Such entities act as referential anchors that underlie the semantics of texts and guide their interpretation. For example, in Europe, by the medieval period, most people were identified simply by a mononym or a single proper name. Family names or surnames began to be expected in the 13th century but in some regions or social classes much later (17th century for the Welsh). Many people shared the same name and the spelling was diverse across vernacular and Latin languages, and also within one language (e.g., Guillelmus, Guillaume, Willelmus, William, Wilhelm). Locations may have disappeared or changed completely, for those that survived well into the 21st century from prehistory (e.g., Scotland, Wales, Spain), they are very ambiguous and also have very different spellings, making it very difficult to identify them [6]. In this article, we focus on exploring temporality in entity detection from historical collections. Thus, we propose a novel technique for injecting additional temporal-aware knowledge by relying on Wikipedia and Wikidata

to provide related context information. More exactly, we retrieve semantically-relevant additional contexts by exploring the time information provided by the historical data collections and include them as mean-pooled representations in our Transformer-based NER model. We consider that adding grammatically correct contexts could improve the error-prone texts due to digitization errors while adding temporality could further be beneficial to handle changes in language or entity names.

The paper is structured as follows: we present the related work and datasets in Sect. 2 and 3 respectively. Our methodology for retrieving additional context through temporal knowledge graphs and how contexts are included within the proposed model is described in Sect. 4. We, then, perform several experiments in regards to the relativity of the time span when selecting additional context and present our findings in Sect. 5. Finally, conclusions and future work are drawn in Sect. 6<sup>3</sup>.

## 2 Related Work

**Named Entity Recognition in Historical Data.** Due to the multiple challenges posed by the quality of digitization or the historical variations of a language, NER in historical and digitized documents is less noticeable in terms of high performance than in modern documents [47, 52]. Recent evaluation campaigns such as the one organized by the *Identifying Historical People, Places, and other Entities* (HIPE) lab at CLEF 2020<sup>4</sup> [16] and 2022<sup>5</sup> [17] proposed tasks of NER and EL in ca. 200 years of historical newspapers written in multiple languages (English, French, German, Finnish and Swedish) and successfully showed that these tasks benefit from the progress in neural-based NLP (specifically driven by the latest advances in Transformer-based pre-trained language models approaches) as a considerable improvement in performance was observed on the historical collections, especially for NER [24, 42, 44].

The authors of [10] present an extensive survey on NER over historical datasets and highlight the challenges that state-of-the-art NER methods applied to historical and noisy inputs need to address. For overcoming the impact of the OCR errors, contextualized embeddings at the character level were utilized to find better representations of out-of-vocabulary words (OOVs) [2]. The contextualized embeddings are learned using language models and allow predicting the next character of strings given previous characters. Moreover, further research showed that the fine-tuning of several Transformer encoders on historical collections could alleviate digitization errors [4]. To deal with the lack of historical resources, [40] proposed to use transfer learning in order to learn models on large contemporary resources and then adapt them to a few corpora of historical nature. Finally, in order to address the spelling variations, some works developed transformation rules to model the diachronic evolution of words and generate a

<sup>3</sup> The code is available at <https://github.com/EmanuelaBoros/clef-hipe-2022-13i>.

<sup>4</sup> <https://impresso.github.io/CLEF-HIPE-2020/>.

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

normalized version processable by existing NER systems [8,23]. While most of these approaches rely generally on the local textual context for detecting entities in such documents, temporal information has generally been disregarded. To the best of our knowledge, several approaches have been proposed for named entity disambiguation by utilizing temporal signatures for entities to reflect the importance of different years [1], and entity linking, such as the usage of time-based filters [26], but not for historical NER.

**Named Entity Recognition with Knowledge Bases.** Considering the complementary behaviors of knowledge-based and neural-based approaches for NER, several studies have explored knowledge-based approaches including different types of symbolic representations (e.g., knowledge bases, static knowledge graphs, gazetteers) and noticed significant improvements in token representations and the detection of entities over modern datasets (e.g., CoNLL [43], OntoNotes 5.0 [35]) [27,43]. Gazetteer knowledge has been integrated into NER models alongside word-level representations through gating mechanisms [31] and Wikipedia has mostly been utilized to increase the semantic representations of possible entities by fine-tuning recent pre-trained language models on the fill-in-the-blank (cloze) task [39,52].

When well-formed text is replaced with short texts containing long-tail entities, symbolic knowledge has also been utilized to increase the contextual information around possible entities [31]. Introducing external contexts into NER systems has been shown to have a positive impact on the entities' identification performance, even with these complications. [48] constructed a knowledge base system based on a local instance of Wikipedia to retrieve relevant documents given a query sentence. The retrieved documents and query sentences, after concatenation, were fed to the NER system. Our proposed methodology could be considered inspired by their work, however, we include the additional contexts at the model level by generating a mean-pooled representation for each context instead of concatenating the contexts with the initial sentence. We consider that having pooled representations for each additional context can reduce the noise that could be created by other entities found in these texts.

**Temporality in Knowledge Graphs.** Recent advances have shown a growing interest in learning representations of entities and relations including time information [7]. Other work [50] proposed a temporal knowledge graph (TKG) embedding model for representing facts involving time intervals by designing the temporal evolution of entity embeddings as rotation in a complex vector space. The entities and the relations were represented as single or dual complex embeddings and temporal changes were the rotations of the entity embeddings in the complex vector space. Since the knowledge graphs change over time in evolving data (e.g., the fact *The President of the United States is Barack Obama* is valid only from 2009 to 2017), A temporal-aware knowledge graph embedding approach [49] was also proposed by moving beyond the complex-valued representations and introducing multivector embeddings from geometric algebras to

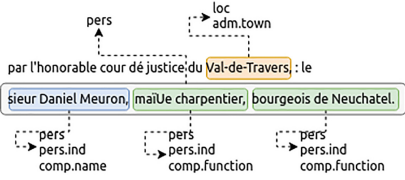


Fig. 1. An example from the hipec-2020 dataset.

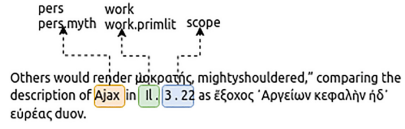


Fig. 2. An example from the ajmc dataset.

model entities, relations, and timestamps for TKGs. Further research [51] presented a graph neural network (GNN) model treating timestamp information as an inherent property of the graph structure with a self-attention mechanism to associate appropriate weights to nodes according to their relevant relations and neighborhood timestamps. Therefore, timestamps are considered properties of links between entities.

TKGs, however, show many inconsistencies and a lack of data quality across various dimensions, including factual accuracy, completeness, and timeliness. In consequence, other research [9] further explores TKGs by targeting the completion of knowledge with accurate but missing information. Moreover, since such TKGs often suffer from incompleteness, the authors of [53] introduced a temporal-aware representation learning model that helps to infer the missing temporal facts by taking interest in facts occurring recurrently and leverage a copy mechanism to identify facts with repetition. The aforementioned methods demonstrate that the usage of TKGs is considered an emerging domain that is being explored, in particular in the field of NLP. The availability of information about the temporal evolution of entities, not only could be a promising solution for improving their semantic knowledge representations but also could provide additional contextual information for efficient NER. To the best of our knowledge, our work is the first attempt to leverage time information provided by TKGs to improve NER.

### 3 Datasets

In this study, we utilize two collections composed of historical newspapers and classical commentaries covering circa 200 years. Recently proposed by the CLEF-HIPE-2022 evaluation campaign [14], we experiment with the hipec-2020 and the Ajax Multi-Commentary (ajmc) datasets.

hipec-2020 includes newspaper articles from Swiss, Luxembourgish, and American newspapers in French, German, and English (19C-20C) and contains 19,848 linked entities as part of the training sets [12, 15, 16]. For each language, the corpus is divided into train, development, and test, with the only exception of English for which only development and test sets were produced [13]. In this case,

**Table 1.** Overview of the `hipe-2020` and `ajmc` datasets. LOC = Location, ORG = Organization, PERS = Person, PROD = Product, TIME = Time, WORK = human work, OBJECT = physical object, and SCOPE = specific portion of work.

	hipe-2020									ajmc								
	French			German			English			French			German			English		
Type	train	dev	test	train	dev	test	train	dev	test	train	dev	test	train	dev	test	train	dev	test
LOC	3,089	774	854	1,740	588	595	-	384	181	15	0	9	31	10	2	39	3	3
ORG	836	159	130	358	164	130	-	118	76	-	-	-	-	-	-	-	-	-
PERS	2,525	679	502	1,166	372	311	-	402	156	577	123	139	620	162	128	618	130	96
PROD	200	49	61	112	49	62	-	33	19	-	-	-	-	-	-	-	-	-
TIME	276	68	53	118	69	49	-	29	17	2	0	3	2	0	0	12	5	3
WORK	-	-	-	-	-	-	-	-	-	378	99	80	321	70	74	467	116	95
OBJECT	-	-	-	-	-	-	-	-	-	10	0	0	6	4	2	3	0	0
SCOPE	-	-	-	-	-	-	-	-	-	639	169	129	758	157	176	684	162	151

we utilized the French and German datasets for training the proposed models in our experimental setup. An example from the French dataset is presented in Fig. 1.

`ajmc` is composed of classical commentaries from the Ajax Multi-Commentary project that includes digitized 19C commentaries published in French, German, and English [41] annotated with both universal and domain-specific named entities (NEs). An example in English is presented in Fig. 2.

These two collections pose several important challenges: the multilingualism (both containing three languages: English, French and German), the code-mixed documents (e.g., commentaries, where Greek is mixed with the language of the commentator), the granularity of annotations and the richness of the texts characterized by a high density of NEs. Both datasets provide different document metadata with different granularity (e.g., language, document type, original source, date) and have different entity tag sets that were built according to different annotation guidelines. Table 1 presents the statistics regarding the number and type of entities in the aforementioned datasets divided according to the training, development, and test sets.

## 4 Temporal Knowledge-based Contexts for Named Entity Recognition

The OCR output contains errors that produce noisy text and complications, similar to those studied by [30]. It has long been observed that adapting NER systems to deal with the OCR noise is more appropriate than adapting NER corpora [11]. Furthermore, [22] showed that applying post-OCR correction algorithms before running NER systems does not often have a positive impact on NER results since post-OCR may degrade clean words during the correction of the noisy ones. To deal with OCR errors, we introduce external grammatically correct contexts into NER systems which have a positive impact on the entity identification performance even in spite of these challenges [48]. Moreover, the

inclusion of such contexts by taking into consideration temporality could further improve the detection of time-sensitive entities. Thus, we propose several settings for including additional context based on Wikidata5m<sup>6</sup> [46], a knowledge graph with five million Wikidata<sup>7</sup> entities which contain entities in the general domain (e.g., celebrities, events, concepts, things) and are aligned to a description that corresponds to the first paragraph of the matching Wikipedia page.

#### 4.1 Temporal Information Integration

A TKG contains time information and facts associated with an entity that provides information about spontaneous changes or smooth temporal transformations of the entity while informing about the relations with other entities. We aggregate temporality into Wikidata5m including the TKG created by [25] and tuned by [18]<sup>8</sup>. This TKG contains over 11 thousand entities, 150 thousand facts, and a temporal scope between the years 508 and 2017. For a given entity, it provides a set of time-related facts describing the interactions of the entity in time. It is thus necessary to combine these facts into a singular element through an aggregation operator over their temporal elements.

We perform a transformation on the temporal information of every fact of an entity in order to combine them into only one piece of temporal information. Let  $e$  be an entity described by the facts:

$$\{F_e\}_{i=1}^n = \{(e, r_1, e_1, t_1), (e, r_2, e_2, t_2), \dots, (e, r_i, e_i, t_i), \dots, (e, r_n, e_n, t_n)\},$$

where a fact  $(e, r_i, e_i, t_i)$  is composed of two entities  $e$  and  $e_i$  that are connected by the relation  $r_i$  and the timestamp  $t_i$ . A timestamp is a discrete point in time which corresponds to a year in this work. The aggregation operator is the function  $AGG \rightarrow t_e$  that takes as input the time information from  $F_e$  and outputs the time information that is associated with  $e$ . Several aggregation operators are possible. Among them, natural options are mean, median, minimum, and maximum operations. The minimum of a set of facts is defined as the oldest fact, and the maximum is the most recent fact. If an entity is associated with four facts spanning over years 1891, 1997, 2006, and 2011, the minimum aggregation operator consists in keeping the oldest, resulting in the year 1891 the time information of the entity. Given that our datasets correspond to documents between 19C and 20C, the minimum operation is more likely to create an appropriate temporal context for the entities. Therefore it is a convenient choice to highlight entities matching the corresponding time period by accentuating older facts. At the end of the aggregation operation 8,176 entities of Wikidata5m are associated with a year comprised between 508 and 2001, filtering out most of the facts occurring during 21C.

<sup>6</sup> <https://deepgraphlearning.github.io/project/wikidata5m>.

<sup>7</sup> <https://www.wikidata.org/>.

<sup>8</sup> <https://github.com/mniepert/mmkb/tree/master/TemporalKGs/wikidata>.

## 4.2 Context Retrieval

Our knowledge base system relies on a local ElasticSearch<sup>9</sup> instance and follows a multilingual semantic similarity matching, which presents an advantage on multilingual querying and is achieved with dense vector field indexes. Thus given a query vector, a  $k$ -nearest neighbor search API retrieves the  $k$  closest vectors returning the corresponding documents as search hits. For each Wikidata5m entity, we create an ElasticSearch entry including an identifier field, a description field and a description embedding field which we obtain with a pre-trained multilingual Sentence-BERT model [37,38]. We build one index on the entity identifier and a dense vector index on the description embedding. We propose two different settings for context retrieval:

- **non-temporal**: This setting uses no temporal information. Given an input sentence during context retrieval, we first obtain the corresponding dense vector representation with the same Sentence-BERT model used during the indexing phase. Then, we query the knowledge base to retrieve the top- $k$  semantically similar entities based on a  $k$ -nearest neighbors algorithm (k-NN) cosine similarity search over the description embedding dense vector index. The context  $C$  is finally composed of  $k$  entity descriptions.
- **temporal- $\delta$** : This setting integrates the temporal information. For each semantically similar entity that is retrieved following **non-temporal**, we apply a filtering operation to keep or discard the entity as part of the context. Given the year  $t_{input}$  linked to the input sentence’s metadata during context retrieval, the entity is kept if its associated year  $t_e$  is inside the interval  $t_{input} - \delta \leq t_e \leq t_{input} + \delta$ , where  $\delta$  is the year interval threshold, otherwise it is rejected. As a result of *AGG*,  $t_e$  results to be the oldest year in the set of facts of entity  $e$  in the TKG. If  $t_e$  is nonexistent,  $e$  is also kept. This operation is repeated until  $|C| = k$ .

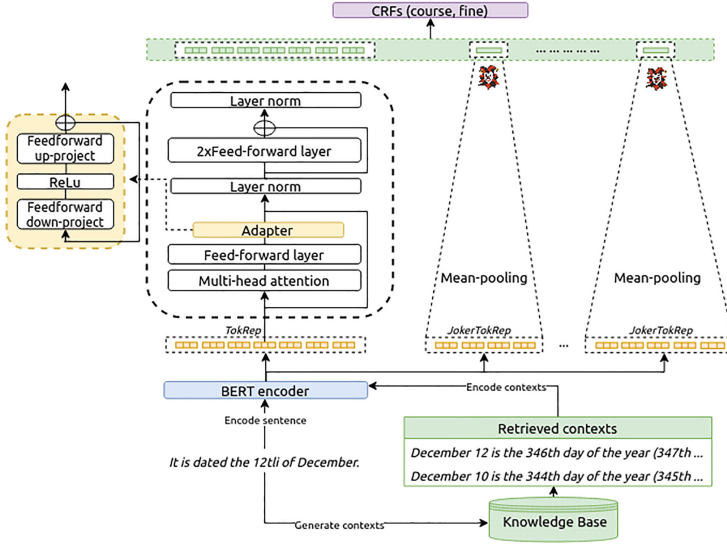
## 4.3 Named Entity Recognition Architecture


*Base Model* Our model consists of a hierarchical, multitask learning approach, with a fine-tuned encoder based on BERT. This model includes an encoder with two Transformer [45] layers with adapter modules [21,34] on top of the BERT pre-trained model. The adapters are added to each Transformer layer after the projection following multi-headed attention and they adapt not only to the task but also to the noisy input which proved to increase the performance of NER in such special conditions [4]. Finally, the prediction layer consists of a conditional random field (CRF) layer.

In detail, let  $\{x_i\}_{i=1}^l$  be a token input sequence consisting of  $l$  words, denoted as  $\{x_i\}_{i=1}^l = \{x_1, x_2, \dots, x_i, \dots, x_l\}$ , where  $x_i$  refers to the  $i$ -th token in the sequence of length  $l$ . We first apply a pre-trained language model as *encoder* for further fine-tuning. The output is  $\{h_i\}_{i=1}^l, H_{[CLS]} = \text{encoder}(\{x_i\}_{i=0}^l)$  where

<sup>9</sup> <https://www.elastic.co/guide/en/elasticsearch/reference/8.1/release-highlights.html>.





**Fig. 3.** NER model architecture with temporal-aware context  s (*context jokers*).

$\{h_i\}_{i=1}^l = [h_1, h_2, \dots, h_i, \dots, h_l]$  is the representation for each  $i$ -th position in  $x$  token sequence and  $h_{[CLS]}$  is the final hidden state vector of  $[CLS]$  as the representation of the whole sequence  $x$ . From now on, we refer to the *Token Representation* as  $TokRep = \{x_i\}_{i=1}^l$  that is the token input sequence consisting of  $l$  words. The additional Transformer encoder contains a number of Transformer layers that takes as input the matrix  $H = \{h_i\}_{i=1}^l \in R_{l \times d}$  where  $d$  is the input dimension (encoder output dimension). A Transformer layer includes a multi-head self-attention  $Head(h): Q^{(h)}, K^{(h)}, V^{(h)} = HW_q^{(h)}, HW_k^{(h)}, HW_v^{(h)}$  and  $MultiHead(H) = [Head^{(1)}, \dots, Head^{(n)}]W_O$ <sup>10</sup> where  $n$  is the number of heads and the superscript  $h$  represents the head index.  $Q_t$  is the query vector of the  $t$ -th token,  $j$  is the token the  $t$ -th token attends.  $K_j$  is the key vector representation of the  $j$ -th token. The *Attn* softmax is along the last dimension.  $MultiHead(H)$  is the concatenation on the last dimension of size  $R^{l \times d}$  where  $d_k$  is the scaling factor  $d_k \times n = d$ .  $W_O$  is a learnable parameter of size  $R^d \times d$ .

By combining the position-wise feed-forward sub-layer and multi-head attention, we obtain a feed-forward layer  $FFN(f(H)) = \max(0, f(H)W_1)W_2$  where  $W_1, W_2$  are learnable parameters and  $\max$  is the *ReLU* activation.  $W_1 \in R^{d \times d_{FF}}$ ,  $W_2 \in R^{d_{FF} \times d}$  are trained projection matrices, and  $d_{FF}$  is a hyper-parameter. The task adapter is applied at this level on  $TokRep$  at each layer and consists of a down-projection  $D \in R^{h \times d}$  where  $h$  is the hidden size of the Transformer model and  $d$  is the dimension of the adapter, also followed by a *ReLU* activation and an up-projection  $U \in R^{d \times h}$ .

<sup>10</sup> We leave out the details that can be consulted in [45].

*Context Jokers* 🃏 For including the additional contexts generated as explained in Sect. 4, we introduce the *context jokers*. Each additional context is passed through the pre-trained *encoder*<sup>11</sup> generating a *JokerTokRep* which is afterwards mean-pooled along the sequence axis. We call these representations *context jokers*. We see them as wild cards unobtrusively inserted in the representation of the current sentence for improving the recognition of the fine-grained entities. However, we also consider that these jokers can affect the results in a way not immediately apparent and can be detrimental to the performance of a NER system. Figure 3 exemplifies the described NER architecture.

## 5 Experimental Setup

Our experimental setup consists of a baseline model and four configurations with different levels of knowledge-based contexts:

- **no-context**: our model as described in Sect. 4.3. In this baseline configuration, no context is added to the input sentence representations.
- **non-temporal**: contexts are generated with the first setting of context retrieval with no temporal information and integrated into the model through *context jokers*.
- **temporal-(50|25|10)**: contexts are generated with the second setting of context retrieval with  $\delta \in \{50, 25, 10\}$  (where  $\delta$  is the time span or year interval threshold) and integrated into the model through *context jokers*.

**Hyperparameters.** In order to have a uniform experimental setting, we chose a BERT-based cased multilingual pre-trained model<sup>12</sup>. We denote the number of layers (i.e., adapter-based Transformer blocks) as  $L$ , the hidden size as  $H$ , and the number of self-attention heads as  $A$ . BERT has  $L=12$ ,  $H=768$  and  $A=12$ . We added two layers with  $H=128$ ,  $A=12$ , and the adapters have  $128 \times 12$  size. The adapters are trained on the task during training. For all context-retrieval configurations, the context size  $|C|$  of an input sentence was set to  $k = 10$ . For indexing the documents in ElasticSearch, we utilized the multilingual pre-trained Sentence-BERT model<sup>13</sup>.

**Evaluation.** The evaluation is performed over coarse-grained NER in terms of precision (P), recall (R), and F-measure (F1) at micro level [12, 28] (i.e., consideration of all true positives, false positives, true negatives and false negatives over all samples) in a strict (exact boundary matching) and a fuzzy boundary matching setting<sup>14</sup>. Coarse-grained NER refers to the identification and categorization of entity mentions according to the high-level entity types listed in Table 1. We refer to these metrics as coarse-strict (**CS**) and coarse-fuzzy (**CF**).

<sup>11</sup> We do not utilize in this case the additional Transformer layers with adapters, since these were specifically proposed for noisy/non-standard text and they do not bring any increase in performance on standard text [4].

<sup>12</sup> <https://huggingface.co/bert-base-multilingual-cased>.

<sup>13</sup> <https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2>.

<sup>14</sup> We utilized the HIPE-scorer <https://github.com/hipe-eval/HIPE-scorer>.

**Table 2.** Results on French, German and English, for the `hipe-2020` and `ajmc` datasets.

	French						German						English					
	hipe-2020			ajmc			hipe-2020			ajmc			hipe-2020			ajmc		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
no-context																		
<i>CS</i>	0.755	0.757	0.756	0.829	0.806	0.817	0.754	0.730	0.742	0.910	0.877	0.893	0.604	0.563	0.583	0.789	0.859	0.823
<i>CF</i>	0.857	0.859	0.858	0.883	0.858	0.870	<b>0.853</b>	0.826	0.839	0.935	0.901	0.917	0.778	0.726	0.751	0.855	0.931	0.891
non-temporal																		
<i>CS</i>	0.762	<b>0.767</b>	<b>0.765</b>	0.829	0.783	0.806	0.759	<b>0.767</b>	<b>0.763</b>	<b>0.930</b>	0.898	0.913	0.565	0.601	0.583	0.828	0.871	0.849
<i>CF</i>	0.862	<b>0.869</b>	0.866	<b>0.906</b>	0.856	0.880	0.847	0.856	0.852	<b>0.949</b>	0.916	<b>0.932</b>	0.741	0.788	0.764	0.885	0.931	0.908
temporal-50																		
<i>CS</i>	<b>0.765</b>	0.765	<b>0.765</b>	0.839	0.822	0.830	0.748	0.756	0.752	0.921	<b>0.911</b>	<b>0.916</b>	<b>0.643</b>	0.617	<b>0.630</b>	0.855	0.882	0.868
<i>CF</i>	<b>0.867</b>	0.867	<b>0.867</b>	0.901	0.883	0.892	0.833	0.842	0.838	0.937	<b>0.927</b>	<b>0.932</b>	<b>0.794</b>	0.762	0.777	0.916	<b>0.945</b>	0.931
temporal-25																		
<i>CS</i>	0.759	0.756	0.757	<b>0.848</b>	<b>0.839</b>	<b>0.844</b>	0.757	0.743	0.750	0.925	0.903	0.914	0.621	0.630	0.625	0.833	0.876	0.854
<i>CF</i>	0.863	0.859	0.861	0.902	<b>0.892</b>	<b>0.897</b>	0.852	0.835	0.843	0.938	0.916	0.927	0.787	0.800	<b>0.793</b>	0.893	0.940	0.916
temporal-10																		
<i>CS</i>	0.762	0.764	0.763	<b>0.848</b>	<b>0.839</b>	<b>0.844</b>	<b>0.760</b>	0.765	0.762	0.917	0.898	0.907	0.605	<b>0.646</b>	0.625	<b>0.866</b>	<b>0.888</b>	<b>0.877</b>
<i>CF</i>	0.863	0.866	0.865	0.902	<b>0.892</b>	<b>0.897</b>	0.852	<b>0.857</b>	<b>0.854</b>	0.936	0.916	0.926	0.760	<b>0.811</b>	0.784	<b>0.922</b>	<b>0.945</b>	<b>0.933</b>
L3i@HIPE-2022																		
<i>CS</i>	<u>0.782</u>	<u>0.827</u>	<u>0.804</u>	0.810	0.842	0.826	<u>0.780</u>	<u>0.787</u>	<u>0.784</u>	<u>0.946</u>	<u>0.921</u>	<u>0.934</u>	0.624	0.617	0.620	0.824	0.876	0.850
<i>CF</i>	<u>0.883</u>	<u>0.933</u>	<u>0.907</u>	0.856	0.889	0.872	<u>0.870</u>	<u>0.878</u>	<u>0.874</u>	<u>0.965</u>	<u>0.940</u>	<u>0.952</u>	0.793	0.784	0.788	0.868	0.922	0.894

## 5.1 Results

Table 2 presents our results in all three languages and datasets (best results in bold). It can be seen that models with additional knowledge-based *context jokers* bring an improvement over the base model with no added contexts. Furthermore, including temporal information outperforms non-temporal contexts. `ajmc` scores show to be higher than `hipe-2020` independently of the language and contexts. We explain this behavior by the small diversity of some entity types of the `ajmc` dataset. For example, the ten most frequent entities from the “person” type represent the 55%, 51.5% and 62.5% from the train, development, and test sets respectively. It also exists an 80% top-10 intersection between train and test sets meaning that eight of the ten most frequent entities are shared between train and test sets. English `hipe-2020` presents the lowest scores compared to French and German independently from the contexts. We attribute this drop in performance to the utilization of the French and German sets during training given the absence of a specific English training set.

The last two rows of Table 2 show the results of our best system [3] during the HIPE-2022 evaluation campaign [15]. This system is similar to the one described in Sect. 4.3 but it stacks, for each language, a language-specific language model and does not include any temporal-aware knowledge. The additional language model motivates the slightly higher results<sup>15</sup>. For half of the datasets, this system outperforms the temporal-aware configurations (underlined values) but with the cost of being language dependent, a drawback that mainly impacts English `hipe-2020` dataset where no training data is available.

<sup>15</sup> We would expect higher results by utilising the temporal information, however, for this experimental setup, we were limited in terms of resources.

**Table 3.** Number of replaced contexts per time span.

	French		German		English	
	train	test	train	test	train	test
<b>temporal-50/25/10</b>						
hipe-2020	120/154/217	42/47/61	325/393/482	12/14/14	192/222/246	77/85/96
ajmc	10/12/12	0/0/0	71/71/73	20/20/20	2/2/2	0/0/0

## 5.2 Impact of Time Intervals

ajmc contains 19th-century commentaries to Greek texts [41] and was created in the context of the Ajax MultiCommentary project<sup>16</sup>, and thus, the French, German and English dataset are about an Ancient Greek tragedy by Sophocles, the Ajax, from the early medieval period<sup>17</sup>. The German ajmc contains commentaries from two years (1853 and 1894), English ajmc, also two years (1881 and 1896), while French ajmc just one year (1886). Due to the size of the collection, hipe-2020 covers a larger range of years. In terms of span, French articles were collected from 1798 to 2018, German articles from 1798 to 1948, and English articles from 1790 to 1960. We, therefore, looked at the difference between the contexts retrieved by the non-temporal and the temporal configurations. Table 3 summarizes these differences for train and test sets and displays the number of contexts that had been filtered and replaced from non-temporal for each time span, i.e.,  $\delta \in \{50, 25, 10\}$ . Overall, the smaller the interval of years, the greater the number of contexts that are replaced. It can be noticed that the number of replaced contexts is smaller for ajmc than for hipe-2020. This is explained by the restrained year span and the lack of entity diversity during these periods. When comparing with the results from Table 2, we can infer that, in general, it is beneficial to implement shorter time intervals such as  $\delta = 10$ . In fact, temporal-10 presents higher F1 scores for ajmc in almost all cases. However, this varies with the language and the year distribution of the dataset.

## 5.3 Impact of Digitization Errors

The ajmc commentaries on classical Greek literature present the typical difficulties of historical OCR. Having complex layouts, often with multiple columns and rows of text, the digitization quality of commentaries could severely impact NER and other downstream tasks like entity linking. Statistically, about 10% of NEs are affected by the OCR in the English and German ajmc datasets and 27.5% of NEs are contaminated in the French corpus. The models with additional context, especially the temporal approaches, contribute to recognizing NEs whether

<sup>16</sup> <https://mromanello.github.io/ajax-multi-commentary/>.

<sup>17</sup> Although the exact date of its first performance is unknown, most scholars date it to relatively early in Sophocles' career (possibly the earliest Sophoclean play still in existence), somewhere between 450 BCE to 430 BCE, possibly around 444 BCE.

contaminated or clean. This contribution is more significant on NEs with digitization errors. It has manifested in a better improvement in recognition of the contaminated NEs compared to the clean ones despite their dominance in the data. In the German corpus, for example, the gain is about 14% points using `temporal-50` compared to the baseline while only 2% points on the clean NEs. Additionally, three-quarters of NEs with 67% of character error rate are correctly recognized whereas the baseline recognized only one-quarter of them. Finally, all the models are completely harmed by error rates that exceed 70% on NEs.

#### 5.4 Limitations

The system ideally requires metadata about the year when the datasets were written or at least a period interval. Otherwise, it will be necessary to use other systems for predicting the year of publication [36]. However, the errors of such systems will be propagated and may impact the NER results.

## 6 Conclusions & Future Work

In this paper, we explore a strategy to inject temporal information into the named entity recognition task on historical collections. In particular, we rely on using semantically-relevant contexts by exploring the time information provided in the collection’s metadata and temporal knowledge graphs. Our proposed models include the contexts as mean-pooled representations in a Transformer-based model. We observed several trends regarding the importance of temporality for historical newspapers and classical commentaries, depending on the time intervals and the digitization error rate. First, our results show that a short time span works better for collections with restrained entity diversity and narrow year intervals, while a longer time span benefits wide year intervals. Second, we also show that our approach performs well in detecting entities affected by digitization errors even to a 67% of character error rate. Finally, we remark that the quality of the retrieved contexts is dependent on the affinity between the historical collection and the knowledge base, thus, in future work, it could be interesting to include temporality information by predicting the year spans of a large set of Wikipedia pages to be used as complementary contexts.

**Acknowledgements.** This work has been supported by the ANNA (2019-1R40226) and TERMITRAD (2020-2019-8510010) projects funded by the Nouvelle-Aquitaine Region, France.

## References

1. Agarwal, P., Strötgen, J., del Corro, L., Hoffart, J., Weikum, G.: diaNED: Time-aware named entity disambiguation for diachronic corpora. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Short Papers). Assoc. Comput. Linguist. Melbourne, Australia, **2**, pp. 686–693 (Jul 2018). <https://doi.org/10.18653/v1/P18-2109>. <https://aclanthology.org/P18-2109>

2. Bircher, S.: Toulouse and Cahors are French Cities, but Ti\*louse and Caa. Qrs as well. Ph.D. thesis, University of Zurich (2019)
3. Boros, E., González-Gallardo, C.E., Giamphy, E., Hamdi, A., Moreno, J.G., Doucet, A.: Knowledge-based contexts for historical named entity recognition linking, pp. 1064–1078. <http://ceur-ws.org/Vol-3180/#paper-84>
4. Boroş, E., Hamdi, A., Pontes, E.L., Cabrera-Diego, L.A., Moreno, J.G., Sidere, N., Doucet, A.: Alleviating digitization errors in named entity recognition for historical documents. In: Proceedings of the 24th conference on computational natural language learning, pp. 431–441 (2020)
5. Boros, E., Nguyen, N.K., Lejeune, G., Doucet, A.: Assessing the impact of OCR noise on multilingual event detection over digitised documents. *Int. J. Digital Libr.*, pp. 1–26 (2022)
6. Boroş, E., et al.: A comparison of sequential and combined approaches for named entity recognition in a corpus of handwritten medieval charters. In: 2020 17th International conference on frontiers in handwriting recognition (ICFHR), pp. 79–84. IEEE (2020)
7. Cai, B., Xiang, Y., Gao, L., Zhang, H., Li, Y., Li, J.: Temporal knowledge graph completion: a survey. arXiv preprint [arXiv:2201.08236](https://arxiv.org/abs/2201.08236) (2022)
8. Díez Platas, M.L., Ros Munoz, S., González-Blanco, E., Ruiz Fabo, P., Alvarez Mellado, E.: Medieval spanish (12th-15th centuries) named entity recognition and attribute annotation system based on contextual information. *J. Assoc. Inf. Sci. Technol.* **72**(2), 224–238 (2021)
9. Dikeoulias, I., Amin, S., Neumann, G.: Temporal knowledge graph reasoning with low-rank and model-agnostic representations. arXiv preprint [arXiv:2204.04783](https://arxiv.org/abs/2204.04783) (2022)
10. Ehrmann, M., Hamdi, A., Linhares Pontes, E., Romanello, M., Douvet, A.: A Survey of Named Entity Recognition and Classification in Historical Documents. *ACM Comput. Surv.* (2022). <https://arxiv.org/abs/2109.11406>
11. Ehrmann, M., Hamdi, A., Pontes, E.L., Romanello, M., Doucet, A.: Named entity recognition and classification on historical documents: a survey. arXiv preprint [arXiv:2109.11406](https://arxiv.org/abs/2109.11406) (2021)
12. 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)
13. Ehrmann, M., Romanello, M., Clematide, S., Ströbel, P.B., Barman, R.: Language resources for historical newspapers. the impresso collection. In: Proceedings of The 12th Language Resources and Evaluation Conference, pp. 958–968 (2020)
14. Ehrmann, M., Romanello, M., Doucet, A., Clematide, S.: Introducing the HIPE 2022 shared task: Named entity recognition and linking in multilingual historical documents. In: *European Conference on Information Retrieval*, pp. 347–354 (2022). [https://doi.org/10.1007/978-3-030-99739-7\\_44](https://doi.org/10.1007/978-3-030-99739-7_44)
15. Ehrmann, M., Romanello, M., Flückiger, A., Clematide, S.: Extended overview of clef HIPE 2020: named entity processing on historical newspapers. In: *CEUR Workshop Proceedings*. 2696, CEUR-WS (2020)
16. Ehrmann, M., Romanello, M., Flückiger, A., Clematide, S.: Overview of CLEF HIPE 2020: Named Entity Recognition and Linking on Historical Newspapers. In: Arampatzis, A., et al. (eds.) *CLEF 2020. LNCS*, vol. 12260, pp. 288–310. Springer, Cham (2020). [https://doi.org/10.1007/978-3-030-58219-7\\_21](https://doi.org/10.1007/978-3-030-58219-7_21)

17. Bellot, P., et al. (eds.): CLEF 2018. LNCS, vol. 11018. Springer, Cham (2018). <https://doi.org/10.1007/978-3-319-98932-7>
18. García-Durán, A., Dumančić, S., Niepert, M.: Learning sequence encoders for temporal knowledge graph completion. arXiv preprint [arXiv:1809.03202](https://arxiv.org/abs/1809.03202) (2018)
19. Hamdi, A., et al.: A multilingual dataset for named entity recognition, entity linking and stance detection in historical newspapers. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2328–2334 (2021)
20. Hamdi, A., Pontes, E.L., Sidere, N., Coustaty, M., Doucet, A.: In-depth analysis of the impact of OCR errors on named entity recognition and linking. *Nat. Lang. Eng.*, pp. 1–24 (2022)
21. Houluby, N., et al.: Parameter-efficient transfer learning for NLP. In: International Conference on Machine Learning, pp. 2790–2799. PMLR (2019)
22. Huynh, V., Hamdi, A., Doucet, A.: When to Use OCR Post-correction for Named Entity Recognition? In: Ishita, E., Pang, N., Zhou, L. (eds.) ICADL 2020. LNCS, vol. 12504, pp. 33–42. Springer, Cham (2020). [https://doi.org/10.1007/978-3-030-64452-9\\_3](https://doi.org/10.1007/978-3-030-64452-9_3)
23. Kogkitsidou, E., Gambette, P.: Normalisation of 16th and 17th century texts in french and geographical named entity recognition. In: Proceedings of the 4th ACM SIGSPATIAL Workshop on Geospatial Humanities, pp. 28–34 (2020)
24. Kristanti, T., Romary, L.: Delft and entity-fishing: Tools for clef HIPE 2020 shared task. In: CLEF 2020-Conference and Labs of the Evaluation Forum. 2696. CEUR (2020)
25. Leblay, J., Chekol, M.W.: Deriving validity time in knowledge graph. In: Companion Proceedings of the The Web Conference 2018, pp. 1771–1776 (2018)
26. Linhares Pontes, E., Cabrera-Diego, L.A., Moreno, J.G., Boros, E., Hamdi, A., Doucet, A., Sidere, N., Coustaty, M.: Melhissa: a multilingual entity linking architecture for historical press articles. *Int. J. Digit. Libr.* **23**(2), 133–160 (2022)
27. Liu, T., Yao, J.G., Lin, C.Y.: Towards improving neural named entity recognition with gazetteers. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 5301–5307 (2019)
28. Makhoul, J., Kubala, F., Schwartz, R., Weischedel, R., et al.: Performance measures for information extraction. In: Proceedings of DARPA broadcast news workshop, pp. 249–252. Herndon, VA (1999)
29. Manjavacas, E., Fonteyn, L.: Adapting vs pre-training language models for historical languages (2022)
30. Mayhew, S., Tsygankova, T., Roth, D.: ner and pos when nothing is capitalized. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) Association for Computational Linguistics, Hong Kong, China, pp. 6256–6261 (Nov 2019). <https://doi.org/10.18653/v1/D19-1650>. <https://aclanthology.org/D19-1650>
31. Meng, T., Fang, A., Rokhlenko, O., Malmasi, S.: Gemnet: Effective gated gazetteer representations for recognizing complex entities in low-context input. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics. Human Language Technol., pp. 1499–1512 (2021)
32. Nguyen, N.K., Boros, E., Lejeune, G., Doucet, A.: Impact analysis of document digitization on event extraction. In: 4th workshop on natural language for artificial intelligence (NL4AI 2020) co-located with the 19th international conference of the Italian Association for artificial intelligence (AI\* IA 2020). 2735, pp. 17–28 (2020)



33. Oberbichler, S., Boros, E., Doucet, A., Marjanen, J., Pfanzelter, E., Rautiainen, J., Toivonen, H., Tolonen, M.: Integrated interdisciplinary workflows for research on historical newspapers: Perspectives from humanities scholars, computer scientists, and librarians. *J. Assoc. Inf. Sci. Technol.* **73**(2), 225–239 (2022)
34. Pfeiffer, J., Vulić, I., Gurevych, I., Ruder, S.: MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, pp. 7654–7673 (2020). <https://doi.org/10.18653/v1/2020.emnlp-main.617>. <https://aclanthology.org/2020.emnlp-main.617>
35. Pradhan, S., Moschitti, A., Xue, N., Ng, H.T., Björkelund, A., Uryupina, O., Zhang, Y., Zhong, Z.: Towards robust linguistic analysis using OntoNotes. In: *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*. Assoc. Comput. Linguist. Sofia, Bulgaria, pp. 143–152. (2013). <https://aclanthology.org/W13-3516>
36. Rastas, I., et al.: Explainable publication year prediction of eighteenth century texts with the bert model. In: *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pp. 68–77 (2022)
37. Reimers, N., Gurevych, I.: Sentence-BERT: sentence embeddings using Siamese BERT-networks. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Assoc. Comput. Linguist. Hong Kong, China, pp. 3982–3992 (2019). <https://doi.org/10.18653/v1/D19-1410>. <https://aclanthology.org/D19-1410>
38. Reimers, N., Gurevych, I.: Making monolingual sentence embeddings multilingual using knowledge distillation. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4512–4525 (2020)
39. Ri, R., Yamada, I., Tsuruoka, Y.: mLUKE: The power of entity representations in multilingual pretrained language models. In: *ACL 2022 (to appear)* (2022)
40. Riedl, M., Padó, S.: A named entity recognition shootout for german. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Short Papers)* **2**, pp. 120–125 (2018)
41. Romanello, M., Najem-Meyer, S., Robertson, B.: Optical character recognition of 19th century classical commentaries: the current state of affairs. In: *The 6th International Workshop on Historical Document Imaging and Processing*, pp. 1–6 (2021)
42. Suárez, P.J.O., Dupont, Y., Lejeune, G., Tian, T.: Sinner clef-hipe2020: sinful adaptation of SOTA models for named entity recognition in French and German. In: *CLEF (Working Notes)* (2020)
43. 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>
44. Todorov, K., Colavizza, G.: Transfer learning for named entity recognition in historical corpora. In: *CLEF (Working Notes)* (2020)
45. Vaswani, A., et al.: Attention is all you need. *Adv. Neural Inf. Proc. Syst.* **30** (2017)
46. Wang, X., Gao, T., Zhu, Z., Zhang, Z., Liu, Z., Li, J., Tang, J.: Kepler: A unified model for knowledge embedding and pre-trained language representation. *Trans. Assoc. Comput. Linguist.* **9**, 176–194 (2021)
47. Wang, X., et al.: Automated concatenation of embeddings for structured prediction. *arXiv preprint arXiv:2010.05006* (2020)



48. Wang, X., et al.: Damo-nlp at semeval-2022 task 11: a knowledge-based system for multilingual named entity recognition. arXiv preprint [arXiv:2203.00545](https://arxiv.org/abs/2203.00545) (2022)
49. Xu, C., Chen, Y.Y., Nayyeri, M., Lehmann, J.: Temporal knowledge graph completion using a linear temporal regularizer and multivector embeddings. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Assoc. Comput. Linguist, pp. 2569–2578 (2021). <https://doi.org/10.18653/v1/2021.naacl-main.202>  
<https://aclanthology.org/2021.naacl-main.202>
50. Xu, C., Nayyeri, M., Alkhoury, F., Shariat Yazdi, H., Lehmann, J.: TeRo: A time-aware knowledge graph embedding via temporal rotation. In: Proceedings of the 28th International Conference on Computational Linguistics. Int. Committee Comput. Linguist. Barcelona, Spain, pp. 1583–1593 (2020). <https://doi.org/10.18653/v1/2020.coling-main.139>  
<https://aclanthology.org/2020.coling-main.139>
51. Xu, C., Su, F., Lehmann, J.: Time-aware relational graph attention network for temporal knowledge graph embeddings (2021)
52. Yamada, I., Asai, A., Shindo, H., Takeda, H., Matsumoto, Y.: LUKE: Deep contextualized entity representations with entity-aware self-attention. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) Assoc. Comput. Linguist, pp. 6442–6454 (2020). <https://doi.org/10.18653/v1/2020.emnlp-main.523>, <https://aclanthology.org/2020.emnlp-main.523>
53. Zhu, C., Chen, M., Fan, C., Cheng, G., Zhang, Y.: Learning from history: modeling temporal knowledge graphs with sequential copy-generation networks. In: Proceedings of the AAAI Conference on Artificial Intelligence. **35**, pp. 4732–4740 (2021)