






Experimenting with Unsupervised Multilingual Event Detection in Historical Newspapers

Emanuela Boros¹(✉) , Luis Adrián Cabrera-Diego^{1,2} ,
and Antoine Doucet¹ 

¹ University of La Rochelle, L3i, F-17000 La Rochelle, France
{[emanuela.boros](mailto:emanuela.boros@univ-lr.fr), [antoine.doucet](mailto:antoine.doucet@univ-lr.fr)}@univ-lr.fr

² Jus Mundi, F-75008 Paris, France
a.cabrera@jusmundi.com

Abstract. To prevent historical knowledge's fading, research in event detection could facilitate access to digitized collections. In this paper, we propose a method for annotating multilingual historical documents for event detection in an unsupervised manner by leveraging entities and semantic notions of event types. We automatically annotate the documents by relying on dependency parse trees and automatic semantic mapping to event-based frames, with a focus on the multilingual transfer between frames and candidate events. The documents are afterward verified by native speakers, Digital Humanities researchers. We also report on experimental results of event detection in historical newspapers with a state-of-the-art model. We demonstrate that our approach allows for easy language adaptation by presenting two study cases with knowledge extracted from German newspapers from 1911 to 1933 regarding events surrounding International Women's Day and from French newspapers between 1900 and 1944 related to the abolition of guillotine executions in France. Our preliminary findings show that this type of approach could alleviate the need for manual annotation by also providing a practical course of action toward unsupervised event detection from multilingual digitized and historical documents.

Keywords: Event detection · Named entity recognition · Historical documents · Historical newspapers

1 Introduction

The digitization of newspapers has greatly improved accessibility and clearly changed the nature of historical research, by enabling easier data access and analysis at scale through multilingual semantic data enrichment [6, 7, 10, 42]. Through better document analysis results and semantic enrichment e.g., named entity recognition (NER), relation extraction (RE), event extraction (EE), the quality of the newspaper data offered by the libraries to its users is substantially improved [12–15]. Preserving the historical memory of entities and events

from historical documents and making them accessible to a larger audience, not only limited to humanities scholars and experts, could lead to better organization of our historical knowledge [2, 38, 44]. Following this statement, this process can be viewed as an area where the detection of events in historical documents can contribute to the construction of more nuanced knowledge bases that could enable further data exploration and help to shape the humanities and historians’ research [38]. Extracting event information from text documents into a structured knowledge base or ontology enables several technologies. For example, text summarization might benefit from the selection of one or more events to yield the best summary with the least extraneous information [18, 30]. Question answering can take advantage of the detected events and they will be able to answer queries about types of events (wars, disease outbreaks, political movements, climate catastrophes, terrorist attacks, etc.) [29, 43].

Therefore, for enabling the development and evaluation of event detection in historical documents, benchmarking plays an important role. However, most of the current datasets in event detection (i.e., MUC [19], ACE 2005 [48]) are not suitable for several reasons, including the high cost of manual annotation of historical texts and the difficulty in defining an event [45]. Besides, different studies have explored how different natural language processing (NLP) tasks, such as named entity recognition (NER) [3, 20, 34, 41] and entity linking (EL) [35, 46], can be impacted by the digitization process. However, to the best of our knowledge, there are no previous works regarding this type of analysis for the event detection task, mostly because there is no data.

Thus, in this paper, we develop a method to automatically discover a set of distinct, salient events from historical newspapers. This is done by leveraging the semantic similarity of contextual representations for detecting event triggers in an unsupervised manner that can be easily adaptable to other languages. The detected events can be used then to speed up the manual annotation or validation of historical corpora.

2 Event Detection in Modern and Historical Datasets

Event Detection in Modern Datasets. Prior work in event detection can be divided in: pattern-based systems [39, 40, 50], machine learning systems based on engineered features (i.e. feature-based) [8, 21, 24, 27], and neural-based approaches [9, 17, 36, 37]. There also has been a lot of interest in approaching this task with external resource-based models which are either feature-based [28, 31] or neural-based [32] combined with resources such as FrameNet [1] which is a linguistic corpus that defines complete semantic frames and frame-to-frame relations, or event data generation as in [22, 49, 51]. The approach proposed in [31] used a probabilistic soft logic (PSL) based approach and a neural network by also leveraging FrameNet to alleviate the data sparseness problem of event detection based on the observation that frames in FrameNet are analogous to events. The authors of [33] also consider that arguments provide significant clues to this task, and adopt a supervised attention mechanism to exploit argument information explicitly for event detection, while also using events from FrameNet, as extra

training data. The model described in [28] also leverages FrameNet by tackling the challenge of the annotation cost and data scarcity by considering that ACE 2005 dataset defines limited and specific event schemes based on FrameNet by expressing event information with frame and building a hierarchy of event schemas that are more fine-grained and have much wider coverage than ACE.

Event Detection in Historical Datasets. When it comes to historical and digitized documents, models rely more on external resources, such as FrameNet and WordNet [16], than on event detection approaches used in the state of the art for modern datasets. FrameNet, for instance, has been highly investigated for event detection in historical and digitized, mostly due to the lack of annotated data. A project proposed in 2004 [25] involved the enhancement of materials drawn from the *Franklin D. Roosevelt Library and Digital Archives* and enabled data exploitation for providing a deeper search and access methods for historians of World War II. The documents were scanned, hand-validated, and enriched with various entities such as person names, dates, locations, and job titles. The work focused on the identification of communicative events in the Memorandum of conversation and implied the extraction of verbs associated with any of the FrameNet “Communication” frames and this communicative event utilized a scheme that assigned the role of communicator to a tagged person or pronoun preceding the verb. Another historical event detection module was proposed to be used for museum collections [10], allowing users to search for exhibits related to particular historical events or actors within time periods and geographic areas, extracted from Dutch historical archives. The authors focused on event detection from manually tagged textual data about the Srebrenica Massacre (July 1995). They specified event triplets and Wordnet concepts denoting event actions, participants, and locations or time markers and identified the historical events through recognition of historical actions. A novel FrameNet-based method was also proposed for performing a computational analysis of Italian war bulletins in World War I and II [7] that had never been digitized before. The bulletins were annotated with different types of information, such as named entities, events, participants, time, and georeferenced locations. Instances of major event types (e.g., bombing, sinking, battles) were established before applying the FrameNet mapping [25].

3 Data Collection

For the data collection and our experiments, we utilized the NewsEye collection¹ that consists of a large selection of European newspapers (1850–1950) in several languages that have been digitized and made available online. The difficulty of detecting events in the NewsEye dataset does not only refer to the automatic text recognition (ATR) or digitization errors, but also to the lack of annotated data in a multilingual setting.

Thus, we decided to annotate two subsets of documents in two low-resource languages, German and French, and to experiment with a state-of-the-art event

¹ <https://www.newseye.eu/>.

detection system in a domain and language adaptation scenario. The documents were collected using the NewsEye platform [26], and annotated by the Digital Humanities groups (native speakers) from the NewsEye consortium, University of Innsbruck (UIBK-ICH), Austria, and the Paul Valéry University Montpellier 3, France. The subjects of the datasets were selected by the annotators, depending on their line of research and interests.

4 Unsupervised Data Annotation

Following the recommendation of Sprugnoli, [45], in this work, we defined an event to be consistent with ACE 2005 [48] and chose the event types and subtypes according to their annotation guidelines². We then automatically assigned a frame category to each event type by consulting the English FrameNet database. FrameNet, as indicated in Sect. 2, is a linguistic corpus containing considerable information about lexical and predicate-argument semantics in the form of frames. A frame, in FrameNet, is defined as a triplet composed of a name, like *Execution*, a set of Frame Elements (FEs), and a list of Lexical Units (LUs).³ An LU is a word or phrase that evokes the corresponding frame, such as *executioner* and *guillotine*. FEs indicate a set of semantic roles associated with the frame, such as *reason*, *instrument* or *place*. Most frames contain a set of exemplars with annotated LUs and FEs.

For linking ACE 2005 event subtypes to FrameNet frames, we start by processing the corpus by extracting all the verbs of the corpus and grouping them using WordNet [16] synsets. Then, the grouped verbs are matched to FrameNet lexical units (LU). Finally, we associate different ACE 2005 event subtypes⁴ to FrameNet by matching frames names to the event subtype names as in [28]. In summary, ACE 2005 event subtypes, are linked indirectly to FrameNet lexical units (LU), which in turn can be seen as event triggers.

For the creation of candidate event mentions, we generate dependency parse trees for each sentence in the dataset⁵. Next, we focus on the extraction of noun-phrases (NPs) that can be pronouns, proper nouns, or nouns, potentially bound with other tokens that act as modifiers, e.g., adjectives or other nouns, that are generally subjects (*nsubj*) or objects (*obj*) (complements of prepositions). Finally, we obtain a triplet composed of the tree *root*, which is generally the verb of the sentence, and its dependents, the *nsubj* and the *obj*. A candidate event mention is, thus, represented by a triplet, where the *root* is commonly a verb, which can possibly be mapped to a lexical unit (LU), similarly as we did for the event trigger candidates. In Fig. 1, we present the dependency parse tree of a sentence in French.

² <https://www ldc.upenn.edu/sites/www ldc.upenn.edu/files/english-events-guidelines-v5.4.3.pdf>.

³ An example of the *Execution* frame can be viewed at Framenet2 website.

⁴ We chose movements, conflictual events, and membership in organizations.

⁵ We used spaCy 3.1+ [23] with the model `xx_ent_wiki_sm` <https://spacy.io/models/xx>.

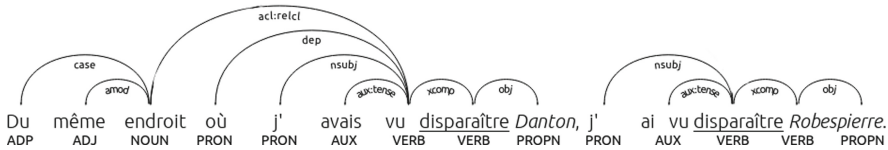


Fig. 1. Example of the correspondence between syntactic arguments of the verbs and participants of the event denoted by the verb (Translation: *From the same place where I had seen Danton disappear, I saw Robespierre disappear.*)

For example, in Fig. 1, there are two triplets both with “disparaître” (to disappear) as root. Specifically, the first triplet is composed of *j’* as subject (*subj*) and “Danton” as object (*obj*). The second triplet has for subject (*subj*) *j’* and for object (*obj*) “Robespierre”. In both triplets, “Danton” and “Robespierre”, besides being objects, they are entities of type person.

For linking candidate event mentions to ACE 2005 event subtypes, we make use of multilingual BERT [11]. To be precise, we use BERT to obtain the contextual representation x of each token in every sentence in the corpus having at least one candidate event mention; $X = [x_0, x_1 \dots x_n]$ where n is the sentence length. Then, from X , we isolate the contextual representation x_i of the token i that represents the candidate root event mention. For example, the first “disparaître” from Fig. 1.

At the same time, we use as well BERT to obtain the contextual representation of the ACE 2005 event subtype by processing FrameNet lexical units (LU) associated, in step 1 as event triggers, for each explored event subtype. Specifically, for a specific event subtype, we concatenated all its event triggers, i.e. lexical units, to generate a pseudo-sentence that is processed by BERT. Then BERT outputs a contextual embedding of the pseudo-sentence, which represents the event subtype.⁶

Finally, in order to consider a candidate event mention, we compare, through cosine similarity, the contextual embedding of an ACE 2005 subtype, with the one of the root of the event mention candidate. If the obtained cosine similarity is greater than 0.7, then the mention candidate is considered to belong to the analyzed ACE 2005 subtype.

For example, for the *Attack* event type in French, we compared the extracted roots with the following set of lexical units that was retrieved from FrameNet: *attack, assault, strike, ambush, assail, raid, bomb, bombing, raid, infiltrate, hit, fire, small, take up arms, fire, airstrike, bombardment, counter-attack, counter-offensive*. After analyzing the results, we observed that two separate sets of event triggers were extracted: (1) known events: *foudroyer* (strike down), *armer* (take up arms), *attaquer* (attack), *frapper* (strike); (2) unseen events: *arracher* (snatch), *déchiqeter* (tear off), *étouffer* (suffocate), *empoigner* (grab), *trancher* (shred).

⁶ We are aware that BERT was trained for representing true sentences rather than pseudo-sentences. However, we consider that BERT might generate an embedding that represents the context in which all the event triggers are frequently used.

5 Evaluation

As indicated in Sect. 2, there is no annotated data for historical event detection, and its creation can be expensive. Thus, to evaluate the unsupervised annotation (Sect. 4), we rely on an indirect assessment based on a fine-tuned language model.

Specifically, we train an event detection system by fine-tuning multilingual BERT [11] on English ACE 2005 following the work of Boros et al. [5]. The goal is that through zero-shot⁷, the event detection system will be able to detect a subset of events in the historical corpus (Sect. 3), which would intersect with those found by the unsupervised method (Sect. 4). Ideally, the spans set by the fine-tuned model should match the spans set by the unsupervised annotation, and thus, we will be able to determine precision, recall, and F-score.

As well, following the work of Boros et al. [5], we explore for our evaluation, the fine-tuning of multilingual BERT on English ACE 2005 along with entity markers. The use of *entity markers* consists in augmenting the input data with a series of special tokens that include the entity type. For example, the sentence from Fig. 1 becomes *From the same place where I had seen [PER_start] Danton [PER_end] disappear, I saw [PER_start] Robespierre [PER_end] disappear.* To do this, we train beforehand a NER system based on a hierarchical architecture that includes a stack of Transformer layers [47] on top of a BERT encoder (BERT-n × Transformer-CRF). This architecture, described in [4], has proved to be robust against OCR errors. The performance of the NER system on this collection is, in terms of F-score 48.32 for German, and 72.71 for French. Once the NER system has been created, we annotate the historical corpus and add the entity markers to the input of the event detection system.

6 Results and Discussion

We present the results obtained through two study cases defined by researchers from the NewsEye project and then, we discuss these results.

International Women’s Day. For this study case, we selected a subset of 207 German articles that mentioned the keyword “Women’s Day” (“Frauentag”) and “International Women’s Day” (“Internationaler Frauentag”) published between 1911 and 1933 in order to analyze the events organized on or around the International Women’s Day. For this subset, we selected events regarding *gatherings* or *movements*. These are revealed by the *Conflict* event type with the *Demonstrate* and *Attack* subtypes and the *Contact* with *Meet* event subtype.

To understand the meaning of the event types in a deeper analysis, we detail several types in the following paragraphs. *Demonstrate* and *Attack* are subtypes of the *Conflict* event type. An *Attack* event is defined as a violent physical act causing harm or damage. For example, in *Um diesen ersehnten Zustand herbeizuführen, entsenden wir unseren Schwestern in der ganzen Welt unsere Grüße und rufen sie auf, beim internationalen Frauentag mit uns gemeinsam gegen die*

⁷ As we use multilingual BERT, even when the training is English, the model should be able to predict events in other languages in a zero-shot manner.

Fortdauer des Krieges zu demonstrieren.⁸, the triggers are: for *Demonstrate*, demonstrieren, and for *Attack*, Krieges. Thus, there are, in this case, two mentions of different types of events.

Table 1. Evaluation of NewsEye German event detection.

| Type | Subtype | P | R | F1 |
|--|-------------|-------|-------|--------------|
| BERT-multilingual-cased | | | | |
| Conflict | Attack | 33.33 | 4.55 | 8.00 |
| Conflict | Demonstrate | 50.00 | 9.09 | 15.38 |
| Contact | Meet | 50.00 | 15.38 | 23.53 |
| | | 44.44 | 9.65 | 15.63 |
| BERT-multilingual-cased+Entity Type Markers | | | | |
| Conflict | Attack | 27.27 | 42.86 | 33.33 |
| Conflict | Demonstrate | 47.06 | 66.67 | 55.17 |
| Contact | Meet | 83.33 | 38.46 | 52.63 |
| | | 52.55 | 49.33 | 47.04 |

We can observe from Table 1 that in the results for the model that does not utilize entity markers, the performance drops significantly, while their presence increases the scores values.

Death Penalty Abolition. In the 1900s, in France, there were regular debates regarding the abolition of the death penalty. For this study case, we selected a subset of 207 French articles that mentioned “guillotine” (same in French) and “death penalty” (“peine de mort”) published between 1900 and 1944. from the following newspapers: *Le Matin*, *L’œuvre* and *Le Gaulois*. We selected events regarding *life*, through the *Die* event subtype, conflictual events (*Conflict* with the *Attack* subtype), and criminal *Justice* events with *Execute* subtype. Due to the digitization and article separation processes, some articles contained an insignificant amount of tokens, thus, we removed those with less than ten tokens⁹.

The results, summarized in Table 2, reveal the capacity of our approach for extracting events while establishing a strong baseline. However, we notice that the scores are rather imbalanced, favoring precision, which could indicate a close similarity between the chosen event types.

Discussion. It must be stated that the low F-scores (Tables 1 and 2) for certain event subtypes are not unexpected. In the first place, we compare two approaches in an indirect way, one using an unsupervised method, and another using a supervised method but trained on different types of documents and language. As well,

⁸ Translation: *In order to bring about this desired state, we send our greetings to our sisters all over the world and call on them to demonstrate together with us against the continuation of the war on International Women’s Day.*

⁹ This threshold was chosen experimentally after we verified the dismissed articles.

Table 2. Evaluation of NewsEye French event detection.

| Type | Subtype | P | R | F1 |
|--|---------|-------|-------|--------------|
| BERT-multilingual-cased | | | | |
| Conflict | Attack | 13.31 | 18.22 | 15.41 |
| Life | Die | 42.30 | 19.60 | 26.83 |
| Justice | Execute | 40.00 | 10.00 | 16.00 |
| | | 31.87 | 15.94 | 19.40 |
| BERT-multilingual-cased+Entity Type Markers | | | | |
| Conflict | Attack | 20.10 | 18.21 | 19.21 |
| Life | Die | 30.82 | 21.41 | 25.30 |
| Justice | Execute | 100.0 | 15.00 | 26.08 |
| | | 50.30 | 18.20 | 23.53 |

the results of the NER system are not perfect, especially for German, which could affect their performance. Nonetheless, the results presented in Tables 1 and 2, show that there is an intersection between the unsupervised annotations and those predicted by the fine-tuned models. Thus, this can signal that the unsupervised approach presented here could be useful for pre-annotating historical documents before being seen by a human. This could accelerate the creation of actual corpora annotated with events and, in the future, automatize the detection of events through machine learning. However, it is clear that the unsupervised method has some limitations. We need to evaluate how well the semantics could have affected the matching of verbs and lexical units in FrameNet, if these mistakes are, for example, the reasons why certain events were not detected in French, as the low recall shows in Table 2.

7 Conclusions

In this paper, we proposed an unsupervised event detection method for detecting events in historical newspapers by relying on available resources. We also obtained promising preliminary results in event detection from multilingual articles surrounding International Women’s Day in German, and the death penalty abolition, in French. We plan in making the dataset publicly available for enabling further research, while envisioning subsequent work regarding an enhanced list of event types and studies concerning the adaptability to other languages.

Acknowledgments. This work has been supported by the European Union’s Horizon 2020 research and innovation program under grants 770299 (NewsEye) and 825153 (Embeddia). Also, it has been supported by the ANNA (2019-1R40226) and TER-MITRAD (2020-2019-8510010) projects funded by the Nouvelle-Aquitaine Region, France.

References

1. Baker, C.F., Fillmore, C.J., Lowe, J.B.: The Berkeley frameNet project. In: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, vol. 1, pp. 86–90 (1998)
2. Bedi, H., Patil, S., Hingmire, S., Palshikar, G.: Event timeline generation from history textbooks. In: Proceedings of the 4th Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA 2017), pp. 69–77 (2017)
3. Boros, E., et al.: Alleviating digitization errors in named entity recognition for historical documents. In: Proceedings of the 24th Conference on Computational Natural Language Learning, pp. 431–441. Association for Computational Linguistics (2020). <https://doi.org/10.18653/v1/2020.conll-1.35>
4. Boros, E., et al.: Robust named entity recognition and linking on historical multilingual documents. In: Cappellato, L., Eickhoff, C., Ferro, N., Névóel, A. (eds.) CLEF 2020 Working Notes. Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum. CEUR-WS (2020)
5. Boros, E., Moreno, J.G., Doucet, A.: Event detection with entity markers. In: Hiemstra, D., Moens, M.-F., Mothe, J., Perego, R., Potthast, M., Sebastiani, F. (eds.) ECIR 2021. LNCS, vol. 12657, pp. 233–240. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-72240-1_20
6. Boschee, E., Natarajan, P., Weischedel, R.: Automatic extraction of events from open source text for predictive forecasting. In: Subrahmanian, V. (ed.) Handbook of Computational Approaches to Counterterrorism, pp. 51–67. Springer, New York (2013). https://doi.org/10.1007/978-1-4614-5311-6_3
7. Boschetti, F., et al.: Computational analysis of historical documents: an application to Italian war bulletins in World War I and II. In: Workshop on Language resources and technologies for processing and linking historical documents and archives (LRT4HDA 2014), pp. 70–75. ELRA (2014)
8. Bronstein, O., Dagan, I., Li, Q., Ji, H., Frank, A.: Seed-based event trigger labeling: how far can event descriptions get us? In: ACL, vol. 2, pp. 372–376 (2015)
9. Chen, Y., Xu, L., Liu, K., Zeng, D., Zhao, J.: Event extraction via dynamic multi-pooling convolutional neural networks. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, vol. 1, pp. 167–176 (2015)
10. Cybulska, A., Vossen, P.: Historical event extraction from text. In: Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pp. 39–43 (2011)
11. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint [arXiv:1810.04805](https://arxiv.org/abs/1810.04805) (2018)
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., et al. (eds.) ECIR 2020. LNCS, vol. 12036, pp. 524–532. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-45442-5_68
13. Ehrmann, M., Romanello, M., Doucet, A., Clematide, S.: Introducing the HIPE 2022 shared task: named entity recognition and linking in multilingual historical documents. In: Hagen, M., et al. (eds.) ECIR 2022. LNCS, vol. 13186, pp. 347–354. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-99739-7_44

14. 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
15. Ehrmann, M., Romanello, M., Najem-Meyer, S., Doucet, A., Clematide, S.: Overview of HIPE-2022: named entity recognition and linking in multilingual historical documents. In: Barrón-Cedeño, A., et al. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2022. LNCS, vol. 13390. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-13643-6_26
16. Fellbaum, C.: Wordnet. In: Poli, R., Healy, M., Kameas, A. (eds) Theory and Applications of Ontology: Computer Applications, pp. 231–243. Springer, Dordrecht (2010). https://doi.org/10.1007/978-90-481-8847-5_10
17. Feng, X., Huang, L., Tang, D., Ji, H., Qin, B., Liu, T.: A language-independent neural network for event detection. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (vol. 2: Short Papers), vol. 2, pp. 66–71 (2016)
18. Filatova, E., Hatzivassiloglou, V.: Event-based extractive summarization (2004)
19. Grishman, R., Sundheim, B.: Message understanding conference-6: a brief history. In: COLING 1996, pp. 466–471 (1996)
20. Hamdi, A., Jean-Caurant, A., Sidere, N., Coustaty, M., Doucet, A.: An analysis of the performance of named entity recognition over OCRed documents. In: 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL), pp. 333–334. IEEE, Illinois, USA (2019)
21. Hong, Y., Zhang, J., Ma, B., Yao, J., Zhou, G., Zhu, Q.: Using cross-entity inference to improve event extraction. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-vol. 1, pp. 1127–1136. Association for Computational Linguistics (2011)
22. Hong, Y., Zhou, W., Zhang, J., Zhou, G., Zhu, Q.: Self-regulation: employing a generative adversarial network to improve event detection. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (vol. 1: Long Papers), pp. 515–526 (2018)
23. Honnibal, M., Montani, I., Van Landeghem, S., Boyd, A.: spaCy: industrial-strength natural language processing in python (2020). <https://doi.org/10.5281/zenodo.1212303>
24. Huang, R., Riloff, E.: Peeling back the layers: detecting event role fillers in secondary contexts. In: ACL 2011, pp. 1137–1147 (2011)
25. Ide, N., Woolner, D.: Exploiting semantic web technologies for intelligent access to historical documents. In: Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04). European Language Resources Association (ELRA), Lisbon, Portugal (2004). <https://www.lrec-conf.org/proceedings/lrec2004/pdf/248.pdf>
26. Jean-Caurant, A., Doucet, A.: Accessing and investigating large collections of historical newspapers with the NewsEye platform. In: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020, pp. 531–532 (2020)
27. Li, Q., Ji, H., Huang, L.: Joint event extraction via structured prediction with global features. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (vol. 1: Long Papers), pp. 73–82. Association for Computational Linguistics, Sofia, Bulgaria (2013). <https://www.aclweb.org/anthology/P13-1008>
28. Li, W., Cheng, D., He, L., Wang, Y., Jin, X.: Joint event extraction based on hierarchical event schemas from FrameNet. IEEE Access 7, 25001–25015 (2019)

29. Liu, J., Chen, Y., Liu, K., Bi, W., Liu, X.: Event extraction as machine reading comprehension. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1641–1651 (2020)
30. Liu, M., Li, W., Wu, M., Lu, Q.: Extractive summarization based on event term clustering. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, pp. 185–188 (2007)
31. Liu, S., et al.: Leveraging FrameNet to improve automatic event detection (2016)
32. Liu, S., Chen, Y., Liu, K., Zhao, J.: Exploiting argument information to improve event detection via supervised attention mechanisms. In: 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017), pp. 1789–1798. Vancouver, Canada (2017)
33. Liu, S., et al.: Exploiting argument information to improve event detection via supervised attention mechanisms (2017)
34. Miller, D., Boisen, S., Schwartz, R., Stone, R., Weischedel, R.: Named entity extraction from noisy input: speech and OCR. In: Proceedings of the sixth conference on Applied natural language processing, pp. 316–324. Association for Computational Linguistics, Seattle, Washington, USA (2000)
35. Mutuvi, S., Doucet, A., Odeo, M., Jatowt, A.: Evaluating the impact of ocr errors on topic modeling. In: Dobрева, M., Hinze, A., Žumer, M. (eds.) ICADL 2018. LNCS, vol. 11279, pp. 3–14. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-04257-8_1
36. Nguyen, T.H., Cho, K., Grishman, R.: Joint event extraction via recurrent neural networks. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 300–309 (2016)
37. Nguyen, T.H., Grishman, R.: Event detection and domain adaptation with convolutional neural networks. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (vol. 2: Short Papers), pp. 365–371. Association for Computational Linguistics, Beijing, China (2015). <https://doi.org/10.3115/v1/P15-2060>
38. Oberbichler, S., et al.: 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 (2021)
39. Riloff, E.: Automatically generating extraction patterns from untagged text. In: AAAI1996, pp. 1044–1049 (1996)
40. Riloff, E.: An empirical study of automated dictionary construction for information extraction in three domains. *Artif. Intell.* **85**(1), 101–134 (1996)
41. Rodriguez, K.J., Bryant, M., Blanke, T., Luszczynska, M.: Comparison of named entity recognition tools for raw OCR text. In: Jancsary, J. (ed.) 11th Conference on Natural Language Processing, KONVENS 2012, Empirical Methods in Natural Language Processing, 19–21 Sept 2012. Scientific series of the ÖGAI, vol. 5, pp. 410–414. ÖGAI, Wien, Österreich, Vienna, Austria (2012). https://www.oegai.at/konvens2012/proceedings/60_rodriquez12w/
42. Rovera, M., Nanni, F., Ponzetto, S.P.: Event-Based access to historical Italian war memoirs. *J. Comput. Cult. Heritage* **14**(1), 1–23 (2021). <https://doi.org/10.1145/3406210>

43. Saurí, R., Knippen, R., Verhagen, M., Pustejovsky, J.: Evita: a robust event recognizer for QA systems. In: Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, pp. 700–707. Association for Computational Linguistics, Vancouver, British Columbia, Canada (2005). <https://aclanthology.org/H05-1088>
44. Shaw, R.B.: Events and periods as concepts for organizing historical knowledge. University of California, Berkeley (2010)
45. Sprugnoli, R.: Event Detection and Classification for the Digital Humanities, Ph. D. thesis, University of Trento (2018)
46. van Strien, D., Beelen, K., Ardanuy, M.C., Hosseini, K., McGillivray, B., Colavizza, G.: Assessing the impact of OCR quality on downstream NLP tasks. In: ICAART 2020 - Proceedings of the 12th International Conference on Agents and Artificial Intelligence vol. 1, pp. 484–496 (2020)
47. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
48. Walker, C., Stephanie, S., Julie, M., Kazuaki, M.: ACE 2005 multilingual training corpus. Linguistic Data Consortium, Technical report (2005)
49. Yang, S., Feng, D., Qiao, L., Kan, Z., Li, D.: Exploring pre-trained language models for event extraction and generation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 5284–5294 (2019)
50. Yangarber, R., Grishman, R., Tapanainen, P., Huttunen, S.: Automatic acquisition of domain knowledge for information extraction. In: 18th International Conference on Computational Linguistics (COLING 2000), pp. 940–946 (2000)
51. Zhang, T., Ji, H., Sil, A.: Joint entity and event extraction with generative adversarial imitation learning. *Data Intell.* **1**(2), 99–120 (2019)