Assessing the impact of OCR noise on multilingual event detection over digitised documents

Emanuela Boros, Nhu Khoa Nguyen, Gaël Lejeune, Antoine Doucet

International Journal on Digital Libraries (IJDL) manuscript No.
(will be inserted by the editor)

Assessing the Impact of OCR Noise on Multilingual Event Detection over Digitised Documents

Emanuela Boros · Nhu Khoa Nguyen · Gaël Lejeune · Antoine Doucet

Received: date / Accepted: date

Abstract Event detection (ED) is a crucial task for natural language processing (NLP) and it involves the identification of instances of specified types of events in text and their classification into event types. The detection of events from digitised documents could enable historians to gather and combine a large amount of information into an integrated whole, a panoramic interpretation of the past. However, the level of degradation of digitised documents and the quality of the optical character recognition (OCR) tools might hinder the performance of an event detection system. While several studies have been performed in detecting events from historical documents, the transcribed documents needed to be hand-validated which implied a great effort of human expertise and manual labor-intensive work. Thus, in this study, we explore the robustness of two different event detection language-independent models to OCR noise, over two datasets that cover different event types and multiple languages. We aim at analysing their ability to mitigate problems caused by the low quality of the digitised documents and we simulate the existence of transcribed data, synthesised from clean annotated text, by injecting synthetic noise. For creating the noisy synthetic data, we chose to utilise four main types of noise that commonly occur after the digitisation process: *Character Degradation*, *Bleed Through*, *Blur*, and *Phantom Character*. Finally, we conclude that the imbalance of the datasets, the richness of the different annotation styles, and the language characteristics are the most important factors that can influence event detection in digitised documents.

Keywords Information Extraction · Event Detection · Digitised Documents

Emanuela Boros
University of La Rochelle, F-17000, La Rochelle, France
E-mail: emanuela.boros@univ-lr.fr

Nhu Khoa Nguyen
University of La Rochelle, F-17000, La Rochelle, France
E-mail: nhu.nguyen@univ-lr.fr

Gaël Lejeune
Sorbonne University, F-75006, Paris, France
E-mail: gael.lejeune@sorbonne-universite.fr

Antoine Doucet
University of La Rochelle, F-17000, La Rochelle, France
E-mail: antoine.doucet@univ-lr.fr

1 Introduction

Event detection (ED) is a challenging subtask of event extraction (EE) that implies the extraction of specific knowledge from certain incidents from texts. This subtask is focused on obtaining event-related information from texts, and, as commonly defined in the field of IE, involves the detection of events. Thus, it deals with the extraction of critical information regarding an event, that can be represented by a keyword, a phrase, a sentence, or a span of text, which evokes that event. For example, an article can elaborate about a new epidemic outbreak or about the election of a new president, and the events to be detected be represented by the name of the epidemic "Spanish flu" or by the words "election" or "elected", etc.

For instance, according to the ACE 2005 annotation guidelines[1] [65], in the sentence "*The comments came on the same day that a prominent Iraqi called for internationally supervised elections in Iraq.*", an event detection system should be able to recognize the word *elections* as a trigger for the event of type *Elect*.

[1] https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-events-guidelines-v5.4.3.pdf

The detection of events in digitised documents can be considered as a building block of historical knowledge with which historians formulate their system of ideas about the past [10,45,56,17,9]. Extracting event information from text documents into a structured knowledge base or ontology enables several technologies. For example, text summarisation might benefit from the selection of one or more events to yield the best summary with the least extraneous information [22,39]. 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.) [57,19,38].

Unfortunately, the knowledge of these events is progressively fading away, especially among young generations. Thus, preserving the historical memory of these events and making them accessible to a larger audience, not limited to humanities scholars and experts, could lead to better organization of our knowledge of history [58,1].

Therefore, for enabling the development and evaluation of event detection in historical documents, benchmark datasets play an important role. However, most of the current datasets in event extraction (i.e., MUC [24], ACE 2005 [65]) are not suitable for this domain for several reasons, besides the high cost of manual annotation of historical texts. Through various digitisation campaigns spread over decades, the possibly degraded historical documents are being digitised using different optical character (OCR) tools applied to documents digitised with variable image definition, resulting in transcribed text of very heterogeneous quality. Since documents are accessible through their transcribed version, errors due to this imperfect OCR are cascading through all downstream applications. There has for instance been strong recent interest in studying the effect of OCR onto other information extraction (IE) tasks (e.g. named entity recognition [5,8,6,43,26,62], named entity linking [37,36], topic extraction [46]). However, to our knowledge, there is no research that studies this impact on event detection.

In this paper, we distinguish between two different event definitions, and we introduce the evaluation framework based on two datasets. As aforementioned, since there are no available historical or digitised datasets for event detection, we chose the following datasets in order to cover different domains and event annotations.

The first one was created along with the data analysis for information extraction in any language (DAnIEL) system [33]. The other one is the ACE 2005 corpora provided by the automatic content extraction (ACE) evaluation[2] [65]. Both datasets will be utilised for all the following experiments.

The DAnIEL dataset (hereafter Daniel-data) consists of numerous multilingual news articles collected from different press threads in the field of health from Google News, focused on epidemic events.

The ACE 2005 dataset is widely utilised in research [48,47,2,7] and covers the most common event types of national and international news, in three languages (Arabic, English, and Chinese) from a variety of sources selected from broadcast news programs, newspapers, newswire reports, internet sources or transcribed audio. This dataset contains event types from broader domains, such as justice-related events (i.e., parole releases, trial hearings, sentences), conflicts (demonstrations, attacks), etc.

We chose to perform experiments with two datasets for leveraging the digitisation issues on documents with different characteristics in order to analyse the impact on more languages, different types of documents, different types of domains.

Consequently, we present two approaches to event detection, both with the ability to handle multilingual data. The first one is based on the DAnIEL system (hereafter Daniel-sys) which is a discourse-level approach that exploits the global structure of news in a newswire. This approach is designed to overcome the difficulty of language adaptation by its character-based characteristic that uses positions of string occurrences in text. We believe that Daniel-sys is adequate for its ability to handle text in any language and its robustness to noise. It only requires two occurrences of adequate substrings, regardless of the recognition of the rest of the text. Its weakness is that it is tailored for epidemic events, although it should be possible to adapt to other domains. In this paper, we will, therefore, experiment with it over epidemic events to decide on the worthiness of its adaptation to other domains.

The second approach is a more recent neural-based method that takes the advantage of unsupervised learning of word representations. This architecture is based on a convolutional neural network (CNN) applied to a local context, more exactly to a window of text around potential keywords that can represent events (we refer to them as *triggers*). This model automatically learns features from the sequence of tokens (word and/or character) and decides if the middle word of the window of text can trigger an event or not. We chose this model for its ability to learn features automatically, independently of the domain, randomly initialised at first, and trained on the event detection task. Determining its

[2] `https://catalog.ldc.upenn.edu/LDC2006T06`

ability to handle noise is another of the objectives of the present paper.

We analyse both models and both datasets systematically. First, for the DANIEL-DATA dataset, we consider both approaches, DANIEL-SYS and the CNN-based approach. For the ACE 2005 dataset, we consider only the CNN-based approach, since the DANIEL-SYS holds the specificity of being focused only on epidemic events, and thus it cannot be applied to the ACE 2005 dataset since it does not contain this event type. We aim at testing the robustness of the models against noise, their ability to treat highly inflected languages, and misspelled or unseen words, which can be either due to the low quality of text or the spelling variants. For these experiments, we present the evaluation general settings separately. Furthermore, we create synthetic data starting from the initial original datasets in order to study the direct impact of OCR on the performance of both approaches.

The remainder of this paper is organised as follows: Section 2 elaborates on issues of event extraction from machine-readable and digitised documents, Section 3 describes the two datasets utilised for our study, Section 4 presents the two systems. The experimental setup (evaluation metrics, noise effects, hyperparameters) are elaborated in Section 5. The systems are then evaluated and analysed in Section 6 (DANIEL-DATA) and Section 7 (ACE 2005). A discussion of all the results is detailed in Section 8 and, finally, Section 9 concludes this study and hints at future work.

2 Related Work

Event Detection in Modern Documents. The research in event detection and extraction technology has important theoretical significance and wide application value and it has been driven forward by a long history that started with the MUC (Message Understanding Conferences) [24] from 1987 through 1998 under the auspices of the US government (ARPA/DARPA) and continued with the Automatic Content Extraction (ACE) program [18]. ACE initiatives are of central importance to the IE field since they provide a set of corpora that are available to the research community for the evaluation and comparison of IE systems and approaches.

Prior work in event detection in the context of the ACE 2005 dataset can be divided in: pattern-based systems [53,54,67], machine learning systems based on engineered features (i.e. feature-based) [28,27,35,11], neural-based approaches [13,48,47,21,2]. The current state of the art for event detection involves neural network models. Several works [48,13] dealt with the event detection problem with models based on CNNs applied on word embeddings. Further, these models were slightly improved [49,48] by the way CNNs are applied to sentences by taking into account the possibility to have non-consecutive n-grams as basic features instead of continuous n-grams. Other proposed methods were based on bidirectional recurrent neural networks (Bi-RNNs) [47] where the usage of memory matrices was systematically investigated to store the prediction information during the course of labeling sentence features. Further, other works have been proposed [21] based on hybrid neural network models, with CNNs and Bi-RNNs. These models combined different neural networks for benefiting from both models' abilities. [21] develop a hybrid neural network (a CNN and an RNN) to capture both sequence and chunk information from specific contexts and use them to train an event detector for multiple languages without any handcrafted features. Some authors went beyond sentence-level sequential modeling, considering that these methods suffer from low efficiency in capturing very long-range dependencies [20] by proposing an approach that goes beyond sentence level. The authors utilised a document representation obtained from an RNN, which can automatically extract cross-sentence clues.

Character embeddings have also been studied and, to analyse the impact of character-level features, the authors of [3,4] proposed to integrate character embeddings, that can capture morphological and shape information about words, into a convolutional model for event detection [2][3].

In the medical field, there are also a number of empirical works that targeted the application of event extraction for the detection of disease outbreaks. With the same objective as DANIEL-SYS [33], BIOCASTER [14,15] analysed disease-related news reports with the purpose of providing a summary of the epidemics. This model was an ontology-based text mining system that processed web text for the occurrence of disease outbreaks by applying named entity recognition and event detection. The major limitation of BIOCASTER is that it is not publicly available, except for the ontology.

Another similar system was the Identification Tool System (GRITS) [29]. The architecture was based on the term frequency-inverse document frequency (TF-IDF) method, pattern-matching tools, and a binary classifier to predict the presence of an epidemic event (disease name) in the text. The system translates non-

[3] However, even though the reported results were better than the model that we experiment within this study, the CNN-based model in [3] that utilises a wide range of convolutional windows, requires a considerable amount of memory resources and therefore could not be put in practice.

English documents using a free translation platform, which can potentially introduce errors to subsequent analysis steps if the translation is incorrect.

Event Detection in Historical Documents. Ryan Benjamin Shaw [58] argues that *"a historian never develops this understanding 'from scratch' or 'discovers' it in the archives. Instead, he produces it by transforming inherited ideas, which may be concepts taken for granted in his culture or concepts developed by his peers and predecessors."* Following this statement, this process can be viewed as an area where the identification and classification of events 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 [51].

For example, a project proposed in 2004 involved the enhancement of materials drawn from the Franklin D. Roosevelt Library and Digital Archives and undertook the encoding, annotation, and multi-modal linkage of a portion of the collection [30]. Moreover, the authors proposed an enhancement of a Web-based interface that enables data exploitation for providing a deeper search and access methods for historians of the World War II. The documents were scanned, hand-validated, and enriched with various entities (such as person names, dates, locations, job titles), part-of-speech, and chunking information. Since for historical research the identification of a range of events is essential, the paper presents a method based on resources like FrameNet[4]. Considering that they worked in a narrow domain, primarily in the Memoranda of Conversation, the focus was only on the identification of communicative events reported in the documents. Therefore, the method implied the extraction of verbs associated with any of the FrameNet "Communication" frame and frame hierarchy. Finally, a communicative event utilised a scheme that assigned the role of communicator to a tagged person or pronoun preceding the verb, and assumes the event comprises the remainder of the sentence.

This simple method for extracting specific targeted event types continued with a computational analysis of Italian war bulletins in War World I and II [10]. This was considered a novel work since WWII Italian war bulletins had never been digitised before. Moreover, other challenges intervened as the type of language (Italian of the first half of the 20th century) and domain (military) required an intense effort of adaptation of existing NLP tools. Bulletins were automatically annotated with different types of information, such as simple and multi-word terms, named entities, events, participants, time, and georeferenced locations. In this work, instances of major event types (e.g., bombing, sinking, battles) were established before applying the FrameNet-based method [30]. The annotated texts and extracted information were also explored with a dedicated Web interface.

Another historical event extraction module was proposed to be used for museum collections [17], 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 historical event extraction from textual data about the Srebrenica Massacre, which was a recent event (July 1995) with a big impact on the public opinions [16]. They defined the event as a historical event model which consists of four slots: a location slot, time, participant, and an action slot.

Since the analysis of the past can help to understand the present and future events, research in forecasting was also proposed. One particular area of research for predictive models using open source text has been the incorporation of events involving actors of political interest. Forecasting political instability has been a central task in computational social science for decades. Effective prediction of global and local events is essential to counter-terrorist planning: more accurate prediction will enable decision-makers to allocate limited resources in a manner most likely to prove effective [9]. These events can cover a range of interactions that span the spectrum from cooperation (e.g. the United States promising aid to Burma) to conflict (e.g. al-Qaeda representatives blowing up an oil pipeline in Yemen). In this paper, the events are represented as a triple consisting of an event code, a source actor, and a target actor, similar to FrameNet frames. For instance, in the sentence *The U.S. Air Force bombed Taliban camps*, the appropriate event triple would be (Employ aerial weapons, U.S. military, Taliban).

Another paper presented an approach for extracting information from historical war memoirs and turning it into structured knowledge [56]. The authors built a corpus that consisted of 25 books, historical memoirs of Italian partisans from World War II in North-Western Italy. Out of 25 books, 20 have been obtained by manual digitisation from the original printed editions, while the remaining five documents have been acquired through automatic conversion from existing digital editions. Despite the good performance of the employed OCR[5], a subsequent manual cleaning has been necessary. This acquisition with considerable effort resulted in a tex-

[4] https://framenet.icsi.berkeley.edu/fndrupal/

[5] The authors utilised the Adobe Acrobat Pro DC OCR software, version 2015. However, the system has long been outdated.

tual corpus of approximately 1.5 million words and over 95,000 sentences.

An important work presented an effort to gather requirements from domain experts about the linguistic annotation of events in the historical domain [61]. This research suggested that the development of annotation guidelines for the analysis of texts in a specific domain must be carried out jointly with experts. Thus, the event was defined consistent with ACE 2005 [65]. Because historical texts are rather general category-spanning diverse topics and genres, the authors put particular effort into developing a set of semantic classes that offer an exhaustive categorisation of events, avoiding too much granularity for annotation purposes but also ensuring informativeness. This led to the definition of 22 event types described in the same manner as the ones proposed by ACE 2005, guidelines, annotated corpus, pre-trained historical embeddings, [6]. The authors also provided theoretical and practical investigations on the topic of event detection and classification in historical texts.

Recently, a corpus of 19th century African American newspapers for event extraction [32] was introduced for the study of the discourse of slave and non-slave African diaspora rebellions published in the periodical press in this period. However, this paper proved that manually annotating real documents has many drawbacks, which often leads to small reliably annotated datasets (the produced dataset contains 115 documents).

However, while several studies have been performed in detecting events from historical documents, the digitisation process needed to be hand-validated, thus implying the need for massive human expertise and causing labor-intensive work for data interpretation.

This progressive digitisation of historical archives provides new textual resources, and the growing interest raises the question of how to provide to the users an account of the knowledge contained in such collections. Nonetheless, the digitisation of documents poses several challenges that either depend on the quality of the documents or the performance of OCR tools. Different studies have been proposed on other IE tasks, e.g. how the named entity recognition and linking models [43, 5, 55, 26] can be impacted by the digitisation process [62, 46], but, to our knowledge, there are no previous works for this type of analysis for the event detection task.

[6] The authors made available the trained models on GitHub: `https://github.com/dhfbk/Histo`

3 Datasets

In this section, we present two different datasets that cover two different event annotations, multiple event types, domains, and languages.

The first one was specifically developed to evaluate Daniel-sys [33]. It is destined for multilingual epidemic surveillance and contains articles on different press threads in the field of health (Google News) focused on epidemic events from different collected documents in different languages, with events simply defined as disease-location pairs.

The second dataset covers a larger set of predefined events, ACE 2005, which contains documents in several languages for the 2005 Automatic Content Extraction (ACE) evaluation[7], with 8 event types, and 33 subtypes covering the most common events of national and international news (from a variety of sources selected from broadcast news programs, newspapers, newswire reports, internet sources and from transcribed audio).

Next, we present both datasets in detail.

3.1 DAnIEL Dataset (Daniel-data)

The corpus consists of health articles from different news sources from Google News that concentrated on epidemic events[33]. Each document was annotated by native speakers for six different languages (English, French, Greek, Russian, Chinese, and Polish). The annotation consisted in the decision of whether an article has a relevant event or not, and if yes, the specification of the disease name and location of the event, a process similar to other works [17, 9]. Aside from language diversity, the length of each document also deviates considerably from each other, varying from just one short paragraph to an article with complete structure.

A tuple of disease name-location defines a relevant Daniel-data event. In occasional cases, the annotation can include the number of victims affected by the disease, making the event a triplet of disease name-location-victims number. By representing an event this way, the task event detection happens at the document level with the goal to identify articles that contain events that fit the description above and extract the best representation of the event, i.e. single or compound words. Because of the spontaneous and haphazard nature of an epidemic outbreak, there is no pre-defined list of types or subtypes of events, thus the detection process is simplified to the detection of a disease name and a location as an epidemic event. A sample of the data

[7] `https://catalog.ldc.upenn.edu/LDC2006T06`

```
"15962": {
  "annotations": [
    [
      "listeria",
      "U.S.",
      "unknown"
    ]
  ],
  "comment": "",
  "date_collecte": "2012-01-12",
  "langue": "en",
  "path": "doc_en/20120112_www.businessweek.
       com_2a21025f6f4dc13c9eb8ebf3d249f3",
  "url": "http://www.businessweek.com/news/2012-01-10/
       listeria-cantaloupe-outbreak-tied-to-flawed-
       safety-practice.html"
},
```


Fig. 1 Example of an event annotated in DANIEL-DATA.

is presented in Figure 1, where the number of victims is unknown.

A common characteristic of an event detection dataset is the lack of balance in distribution. In the case of this corpus, documents that are relevant to epidemic events only occupy about 10% of the total dataset, which is very sparse. The number of documents per language, however, is relatively balanced with 352 Polish documents (30 relevant), 446 in Chinese (16 relevant), 390 in Greek (26 relevant), and 475 in English (31 relevant). French is the only exception, having five times more documents than the others, with 2,733 documents, in which 340 of them are relevant. In total, the dataset comprises 4,822 documents (489 relevant).

DANIEL-DATA is annotated at document-level, which differentiates it from other datasets used in research for the event detection task. A document is either reporting an event (disease-location pair, and sometimes the number of victims) or not. In order for us to be able to compare the two different models that we proposed, we transformed this annotation to sentence-level. The annotations provided in DANIEL-DATA at document-level are looked up in the corresponding file and the found offsets are attached to them. For example, the article below has the following annotations, at document level: *U.S.* and *listeria*.

Figure 2 presents an example of event detection in an English document.

In this example, in the sentence, *A listeria outbreak that killed 30 people and sickened another 146 may have been avoided if a Colorado cantaloupe processor had followed U.S. guidelines and washed the fruit in chlorinated water, a congressional investigation found. [. . .]*, we are able to annotate listeria at the relative positions to the entire article 2–9. The process is automatic and continues in the same manner as for the other annotations. From a total of 1,268 (disease names, locations, or the number of affected persons), 1,084 were identified in DANIEL-DATA, thus 85.48% of the annotations were correctly found.

Listeria Cantaloupe Outbreak Tied to Flawed Safety Practice

Jan. 10 – A listeria outbreak that killed 30 people and sickened another 146 may have been avoided if a Colorado cantaloupe processor had followed U.S. guidelines and washed the fruit in chlorinated water, a congressional investigation found. Jensen Farms in Granada, Colorado, also added new processing equipment that may have led to contamination, according to the report issued today by the House Energy and Commerce Committee . . .

Fig. 2 Representation of the occurrences of the event components in a relevant English document. The name of the disease and the location are underlined.

3.2 ACE 2005 Dataset

For our experiments, we utilised the ACE 2005 corpus provided by the ACE evaluation. ACE events are restricted to a range of types, each with a set of subtypes. ACE 2005 contains collections of documents in multiple languages (Chinese, Arabic, and English) with various types annotated for entities, relations, and events, from various information sources (e.g., broadcast conversations, broadcast news, and telephone conversations). The data were created by Linguistic Data Consortium (LDC) with support from the ACE Program.

For comparison purposes, for the experimental setup, we utilised the same data split, detailed in Table 1, as in previous studies on this dataset [48,47]. The test set contains 40 newswire articles (672 sentences), the development set comprises 30 other documents (863 sentences) and the training set comprises the remaining 529 documents (14,849 sentences).

Table 1 English ACE 2005 corpus summary, Newswire (NW), Broadcast Conversation (BC), Broadcast News (BN), Telephone Speech (CTS), Usenet Newsgroups (UN), and Weblogs (WL). The number of documents annotated with one or multiple events is reported in brackets.

Total	NW	BN	BC	WL	UN	CTS
599	106	226	60	119	49	39
(553)	(104)	(211)	(60)	(93)	(47)	(38)

The corpus has eight types of events, with 33 subtypes. These are the types of events:

- Business: Start-Org, Merge-Org, End-Org, Declare-Bankruptcy
- Conflict: Attack, Demonstrate
- Contact: Meet, Phone-Write
- Life: Be-Born, Marry, Divorce, Injure, Die
- Movement: Transport
- Justice: Arrest-Jail, Release-Parole, Trial-Hearing, Charge-Indict, Sue, Convict, Sentence, Fine, Execute, Extradite, Acquit, Appeal, Pardon
- Transaction: Transfer-Ownership, Transfer-Money
- Personnel: Start-Position, End-Position, Nominate, Elect

An ACE event is represented by an *event mention* (a text contains an event of a specific type and subtype), an *event trigger* (the word that expresses the event mention), an *event argument* (a participant in the event of a specific type) and an *argument role* (the role that the entity has in the event).

In the context of ACE 2005, the event extraction (EE) task has two sub-tasks:

1. *Event detection*: the detection of the texts that contain events of specific types and the extraction of the event trigger from the text that expresses that type of event; and
2. *Event argument extraction*: the detection of entities and their role in the event.

However, in this study, we only tackle the event detection task. A document can be characterised by multiple events present at the sentence level, or no events at all. If we consider, for instance, this example from the ACE 2005 dataset:

There was the free press in Qatar, Al Jazeera, but its offices in Kabul and Baghdad were bombed by Americans., annotated as in the Figure 3, an event detection system should output:

- *event mention*: this sentence contains an event of type *Conflict* and subtype *Attack*
- *event trigger*: this event of type *Conflict* and subtype *Attack* is triggered by the word bombed

An event argument extraction system should output[8]: the *event arguments*: *Kabul* and *Baghdad*, which are entities of type *location*, and *Americans* which are considered an entity of type *Person*, and their *event argument roles*: *Kabul* and *Baghdad* are *Places* and *Americans* have the *Attacker* role.

[8] We remind here that this sub-task is not treated in this study. Because event detection is already challenging, we base our experiments only on ED.

```
<event TYPE="Conflict" SUBTYPE="Attack">
  <event_mention>
    <ldc_scope>
      <charseq START="3074" END="3181">There was the
      free press in Qatar, Al Jazeera but its' offices
      in Kabul and Baghdad were bombed by Americans.
      </charseq>
    </ldc_scope>
    <anchor>
      <charseq START="3163" END="3168">bombed
      </charseq>
    </anchor>
    <event_mention_argument ROLE="Attacker">
        <charseq START="3173" END="3181">Americans
        </charseq>
    </event_mention_argument>
    <event_mention_argument ROLE="Place">
        <charseq START="3140" END="3144">Kabul
        </charseq>
    </event_mention_argument>
    <event_mention_argument ROLE="Place">
        <charseq START="3150" END="3156">Baghdad
        </charseq>
    </event_mention_argument>
    <event_mention_argument ROLE="Target">
        <charseq START="3124" END="3156">its' offices
        in Kabul and Baghdad.
        </charseq>
    </event_mention_argument>
  </event_mention>
</event>
```

Fig. 3 ACE 2005 event annotation example.

4 Approaches

This section describes the approaches that will be evaluated, DANIEL-SYS and the CNN-based model. The hyperparameters for both models are detailed in Section 5, where the experimental setup is presented, because they are specific for each dataset.

4.1 DANIEL System (DANIEL-SYS)

DANIEL [33] stands for Data Analysis for Information Extraction in any Language and it is an approach at discourse-level, as opposed to the commonly used analysis at sentence-level, by exploiting the global structure of news (repetition of key information at key positions in the text) as defined in [40,41]. Entries in the system are news texts, including the title and the body of text, the name of the source when available, and other metadata (e.g. date of article). As the name implies, the system is capable of working in a multilingual setting due to the fact that it does not make use of any word-based algorithm (like tokenizers), which are highly language-specific, but rather a character-based algorithm that relies on repetition and position [33]. DANIEL-SYS uses

a minimal knowledge base, its central processing chain includes four phases, as shown in Figure 4:

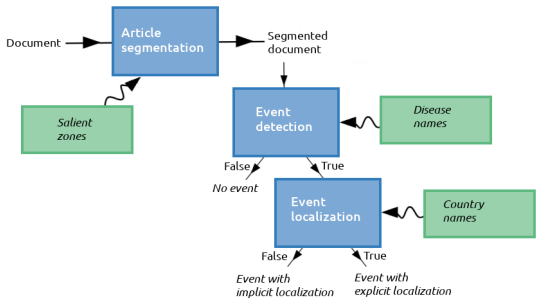

Fig. 4 Event detection pipeline in DANIEL-SYS.

1. *Article segmentation*: The system first divides the document into stylistic segments: title, header, body, and footer. The purpose is to identify salient zones (title, header, footer of the document) where important information is usually repeated.
2. *Pattern extraction*: The system looks for repeated substrings at the salient zones aforementioned and determines whether they are maximal or not. A maximal substring is a string that cannot be extended to either its left or right side [64].
3. *Pattern filtering*: maximal repeated substrings extracted at the previous phase are matched to a list of disease/location names that was constructed by crawling Wikipedia[9]). The reason for using Wikipedia to build the knowledge base is that it allows adding new languages at a minimal cost since there is no need for the assistance of a native speaker and information on Wikipedia can be easily crawled from one language to another. The information about the language of the text can be given in advance or computed with a language identification module.
4. *Detection of disease−location pairs* (in some cases, the number of victims also): The end result of processing a document with DANIEL-SYS is one or more events that are described by disease-location pairs[10].

The document is the main unit of DANIEL-SYS and has language-independent organisational properties, as described in the following paragraph. The assumption is that the document-detectable features at a document granularity offer high robustness at the multilingual scale. The author suggests using the text as a minimal unit of analysis beyond its relation to the genre from which it came. The press article is thus of this type, which has precise rules: the structure of the press article and the vocabulary used are established and there are well-defined communication aims known to the source as well as the target of the documents. These rules, at a higher level than the grammatical rules, are very similar in different languages, and from the knowledge of these rules, remarkable positions are defined which are independent of languages, following previous research related to news genre invariants [23,41].

In the news genre, the different positions in the text are defined here as follows: the beginning of the text (ideally composed of the title of the article), beginning of body (containing the first two paragraphs), end of body (foot) (comprising the last two paragraphs, rest of body (made up of the rest of the textual elements (e.g. paragraphs)).

The fixed structure of a news article can be demonstrated with the following example:

Title: One was an arrogant bully. The other was a nervous wreck. So what is the truth about the war of Van Gogh's ear?
Paragraph 1/38: The iconic story about Dutch-born painter Vincent Van Gogh cutting off his own ear and presenting it as a gift to his favourite prostitute may not be true after all. **Paragraph 2/38**: Or so say some German art historians, who now claim the famous ear was cut off in a fight with rival artist Paul Gauguin.
...
Paragraph 29/38: Gauguin took himself off to Tahiti where he entertained under-age mistresses, consumed vast quantities of absinthe and morphine and died of syphilis in 1903.
...
Paragraph 37/38: Even this did not put an end to his torture. Van Gogh staggered back to the inn where he was lodging and lingered for two days before dying.
Paragraph 38/38: His poignant last words, according to Theo, the distraught brother who had rushed to his side, were: "The sadness will go on for ever."

Fig. 5 Representation of the occurrences of different terms in an English document. The name of the disease, in red can lead to a classification error if analysed at the sentence level, but it is not repeated and thus not considered as a relevant descriptor by DANIEL-SYS. The names of the two painters in question are in blue. The constituents of the event mainly described in the article appear in orange.

In Figure 5, one can see that important pieces of information are repeated at easily identifiable positions in the text. These elements are usually found in at least two of these positions. We can see that the terms Gauguin and Van Gogh have a rich distribution. The same applies to the terms relating to Van Gogh's cut ear. Position and repetition, therefore, make it possible here to prioritise information without resorting to local analysis.

By avoiding grammar analysis and the usage of other NLP toolkits (e.g part-of-speech tagger, dependency parser) and by focusing on the general structure of journalistic writing style [25,41], the system is able to detect crucial information in salient zones that are peculiar to

[9] This process is done by using the interlingual links coming from English infectious diseases Wikipedia pages.
[10] If no location matches the previous rules, the system assumes that the event takes place in the country of the "source" metadata (*Implicit Location Rule* [33]).

this genre of writing: the properties of the journalistic genre, the style universals, form the basis of the analysis.

Moreover, because DANIEL-SYS does not rely on any language-specific grammar analysis, and considers text as sequences of strings instead of words, it can quickly operate on any foreign language and extract crucial information early on and improve the decision-making process. This is pivotal in epidemic surveillance since timeliness is key, and more often than not, initial medical reports are in the vernacular language where patient zero appears [33].

4.2 Convolutional Neural Network-based Model

We chose a convolutional neural network (CNN) based model proposed and explored by [48,2,3], where the event detection (ED) task is modelled as a word classification task with word embeddings as features. A number of CNNs, recurrent neural networks (RNNs), and other neural architectures have been proposed for event detection [13,47,21]. This CNN-based model was one of the first neural network models to be proposed for the event detection task, along with the idea of classifying words with a window of contextual words from which CNN features on different levels are extracted, and it is considered as a reference model for effectively detecting events in a text.

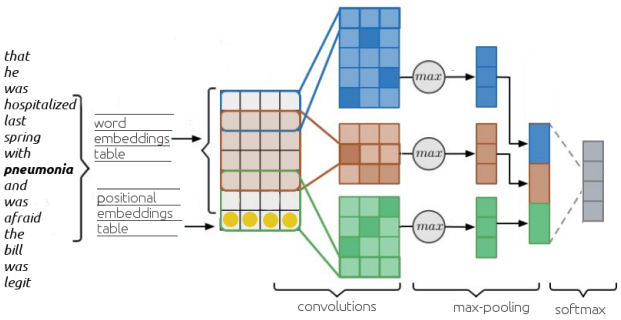

Fig. 6 CNN model for event detection, where *pneumonia* is the current event candidate in a context window of $2 \times 7 + 1$ words, where 7 words surround the event candidate *pneumonia* on its right side and on its left side. Figure from [2].

The methodology of this model is as follows. Considering a sentence, we want to predict, for each word of the sentence, if the current token is a possible trigger of an event, thus the classification type is binary. The current token $x^{(i)}$ is surrounded by a context that constitutes a sliding window over the input text that consists of the input format for the CNN. The maximum size of a sentence is established on the training data. In order to consider a limited-sized context, longer sentences are trimmed and shorter ones are padded with a special token. Let $x = [x^{(0)}, x^{(1)}, ..., x^{(N)}]$ be a sentence with words from 0 to N. Given a document, we first generate a set of event candidates $\mathcal{T}$ (every word in the document). For each event candidate $x^{(i)} \in \mathcal{T}$, we associate it a context window with a size of $2 \times n + 1$, thus a trigger candidate $x^{(0)}$ is represented as $x = [x^{(-n)}, x^{(-n+1)}, \ldots, x^{(0)}, \ldots, x^{(n-1)}, x^{(n)}]$. Each context token $x^{(i)}$, including the event candidate (middle word), has as features the embedding for the word itself and the relative position of the token to the trigger candidate $x^{(0)}$. In this case, the distance 0 will be attributed to the trigger candidate $x^{(0)}$ and $-n$, $+n$ to the marginal tokens of the window, all the other relative distances in between $-n$ and $+n$ belong to the tokens in between. The position of an event trigger can be an informative signal for this prediction task. Each core feature is embedded and represented in a d-dimensional space.

Each feature (word, distance) is mapped to a real-valued vector (embedding). The word features are represented by word embeddings that are either initialised randomly, drawn from a uniform distribution, or by pre-trained word embeddings. For ACE 2005, we utilised the pre-trained English word embeddings *Word2vec* that were trained on Google News [42,2]. Since DANIEL-DATA contains six different languages, we randomly initialised the embeddings in this case. The position features are positional embeddings. To embed the relative distance i of the token $x^{(i)}$ to the current token $x^{(0)}$. The positional embeddings are initialised randomly and they are then trained as regular parameters in the network [48,2].

5 Experimental Setup

In order to circumvent the drawback of lacking ground-truth data with high variability, the objective of our setup is to create appropriate datasets for measuring the impact of noisy input data in subsequent NLP analysis.

First, raw text from both datasets was extracted and converted into clean images. For this step, we employed the ImageMagick[11] tool with a white background, with a width of 1024 and an adjustable height depending on the amount of text in a document, with Arial

[11] `https://imagemagick.org`

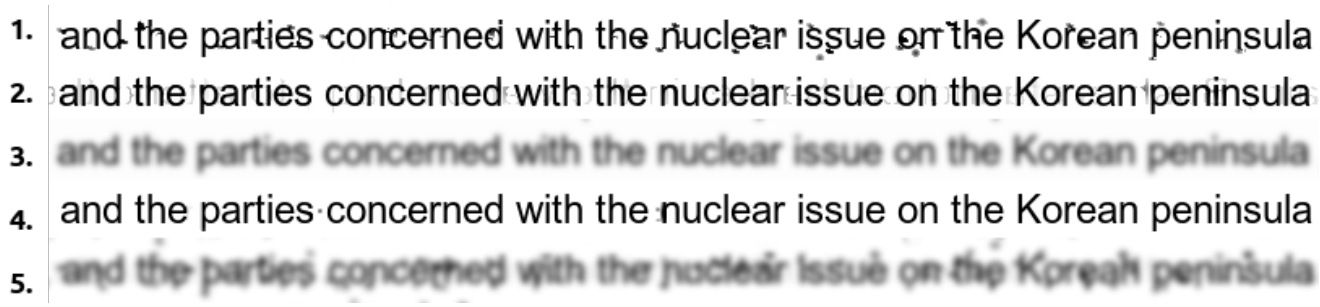

Fig. 7 Example of types of noise applied on ACE 2005 dataset: (1) *Character Degradation*, (2) *Bleed Through*, (3) *Blur*, (4) *Phantom Character*, and (5) all mixed together.

Unicode Regular font[12] of size 24. These parameters were chosen so that all text document sizes can be readable and adaptable to all the different languages in our experiments.

Next, for the simulation of different levels of degradation on these images, we utilised DocCreator [31], which was successfully utilised in previous studies regarding the automatic generation of noise [37,26]. The rationale is to simulate what can be found in deteriorated documents due to time effect, poor printing materials, or inaccurate scanning processes, which are common conditions in historical newspapers. Generally, the output of an OCR tool can contain different types of errors: misspelled characters (substitutions), spurious symbols (insertions), missing characters (deletions). We chose different noise effects in order to imitate such errors and synthetically produce realistic images. A detailed discussion on the reasons for our selection for OCR errors and their types can be found in [50].

After processing the corpus, all the text was extracted from noisy images, for the clean images (without any change) and the noisy synthetic ones. The Tesseract optical character recognition (OCR) Engine v4.0[13] [59] was utilised to produce the digitised documents[14].

5.1 Noise Effects

For creating the noisy synthetic data, we chose to utilise four main types of noise that commonly occur after the digitisation process: *Character Degradation*, *Bleed Through*, *Blur*, and *Phantom Character*. An example with the degradation levels is illustrated in Figure 7. The noise levels were chosen with the highest level of difficulty. Thus, the following values of DocCreator noise types are: *Character Degradation* (2-6), *Phantom Character* (Very Frequent), *Blur* (1-3), *Bleed Through* (80-80).

Character Degradation adds small ink dots on characters to emulate the age effect on articles. Most common character degradations are due to the age of the document itself and the printing or writing processes, such as ink splotches, white specks, or streaks. DocCreator locally degrades the image in the neighbourhood of the boundaries of the characters and then, noise is generated to create some small ink spots near characters or to erase some character's ink area. This effect is visible in the first line of text in Figure 7. By adding these ink dots, *Character Degradation* introduced spurious symbols (insertions). This line of text was recognized by the OCR tool as "and the partiés concerned with the siuclear issue othe Korean peninsula", where, for example, an accent was added to the transcription of "parties", thus becoming "partiés", and "nuclear" was recognized as "siuclear".

Bleed Through appears in double-paged document image scans where the content of the backside appears in the front side as interference. The nonlinear nature of the ink seepage can be seen from the interference patterns of the verso side that show through onto the recto side. Seepage of ink through a paper is a complex physical phenomenon in which many parameters, such as thickness, the characteristics of the paper, the distribution of the paper fibers, and ink quality, are involved. The *Bleed Through* noise, in the second line of text in Figure 7, even if we applied its highest level provided by DocCreator, is slightly visible. *Bleed Through* also inserts spurious symbols, and, in our case, the OCR tool recognized "and the parties concerned with the nucleareissue on the Korean tpeninsula", thus "e" was intro-

[12] `https://docs.microsoft.com/en-us/typography/font-list/arial-unicode-ms`

[13] `https://github.com/tesseract-ocr/tesseract`

[14] We assume that using different major versions of Tesseract (e.g. from 3.x to 4.x) may affect our results since the OCR engine has changed considerably according to the changelog. However, since we chose the last version available, it might be too tedious and time-consuming to perform experiments with different Tesseract versions in light of a different OCR engine.

duced between "nuclear" and "issue", and "peninsula" became "tpeninsula".

Blur is a common degradation effect encountered during a typical digitisation process. The DocCreator blur defect mimics the very slight blur that appears when the scanner is incorrectly set (a large blur is easily detected by scanners). When applying the *Blur* effect, we chose a range from 1 to 3, a rather small range, due to the fact that it considerably affected the quality of the image. This effect can be seen in the third line of text in Figure 7. *Blur* introduced, besides misspelled characters (substitutions), more deletion errors, thus missing and misspelled characters are clearly visible in the following transcription: "and the parties concemed with the nuciear ssue on the Korean peninsula", where "rn" was recognized as "m" in "concerned", "l" became "i" in "nuclear", and the starting letter "i" from "issue" was not detected.

Phantom Character appears when characters erode due to excessive use of documents while being manually printed (using a wooden or metal character). After many uses, a printing character can be eroded. It is thus possible that ink reaches the borders of the piece, and borders are then printed on the sheet of paper. DocCreator provides an algorithm that reproduces such ink apparition around the characters. However, *Phantom Character* is not very visible in Figure 7 (the fourth text line), as it mostly affects the margins of an image, by also introducing spurious symbols. However, in this example, the OCR tool "and the parties·concerned with the :nuclear issue on the Korean peninsula", we observed that noise as in "·" and ":" was detected.

When applying the effects altogether, in the last text line of the Figure 7, an increased amount of OCR errors were introduced. Thus, the transcribed text became "and the parties concerned with the nuciear ssué on the Korealt peninSula", and we notice that the output contained a combination of the different types of errors that an OCR can have: substitutions ("nuciear" instead of "nuclear"), insertions ("Korealt" instead of "Korean"), and deletions ("ssué" instead of "issue").

5.2 General Evaluation Settings

The general evaluation setup has two main settings:

1. Experiments with *original* data: we report the performance scores when the models use the initial original textual datasets (where no OCR and no noise effect has been applied).
2. Experiments with noisy data: we report the performance of the models after the text has been passed through the digitisation process.
   - *Clean*: documents obtained after applying the OCR tool on the original data;
   - *CharDeg*: documents obtained after applying the OCR tool on the original data that has been tempered by the *Character Degradation* noise effect;
   - *Bleed*: documents obtained after applying the OCR tool on the original data that has been tempered by the *Bleed Through* noise effect;
   - *Blur*: documents obtained after applying the OCR tool on the original data that has been tempered by the *Blur* noise effect;
   - *Phantom*: documents obtained after applying the OCR tool on the original data that has been tempered by the *Phantom Character* noise effect;
   - *All Effects*: documents obtained after applying the OCR tool on the original data that has been tempered by the all the previous noise effects.

For the evaluation of the performance of the event detection task, we use the standard metrics: Precision (P), Recall (R), and F-measure (F1). For measuring the document distortion due to the OCR process, we also report the standard metrics: *character error rate* and *word error rate*. The P, R, and F1 are defined by the following equations, where TP: True Positives, FP: False Positives, and FN: False Negatives:

$$P = \frac{TP}{TP + FP} \tag{1}$$

$$R = \frac{TP}{TP + FN} \tag{2}$$

$$F1 = \frac{2PR}{P + R} \tag{3}$$

For all our experiments, we also report the standard deviation (±). This is because for Daniel-data, we use a cross-validation resampling method with five experiments for each setting, and for ACE 2005, we run five experiments with different random seeds, in order to obtain stable results and notice their variance. Note that, in some cases, a standard deviation of 0.00 could happen, depending on the imbalance in the dataset. For example, in the Daniel-data, there is a limited amount of documents containing events (relevant) that depends on the low-resource nature of particular languages. Thus, even with cross-validation sampling, the number of relevant articles remain stable and could be difficult to detect, producing the same results for each data sample.

Character error rate (CER) is defined as:

$$CER = \frac{i_c + s_c + d_c}{n_c} \tag{4}$$

Table 2 Evaluation of the CNN-based model and DANIEL-SYS on the original test data for *event identification*.

		Polish	Chinese	Russian	Greek	French	English	All Languages
DANIEL-SYS	P	80.33±12.26	40.00±54.77	76.28±14.61	100.0±0.0	56.52±5.61	76.33±5.82	64.08±3.71
	R	77.66±13.72	40.00±54.77	83.80±10.26	100.0±0.0	89.03±2.20	100.0±0.0	89.68±1.94
	F1	**78.11**±8.12	40.00±54.77	**79.71**±12.36	**100.0±0.0**	69.03±4.39	**86.47**±3.87	74.71±2.78
CNN	P	84.60±18.45	63.00±10.77	91.64±10.53	80.77±15.77	82.88±0.70	97.88±2.63	82.98±0.74
	R	54.64±3.11	80.00±24.49	68.33±9.34	59.99±7.28	75.83±6.74	61.92±5.62	71.16±5.44
	F1	57.93±5.21	**68.09±14.87**	72.27±5.53	64.26±8.88	**78.37±4.47**	68.29±7.04	**75.34±4.19**

where n_c is the ground-truth in terms of characters, i_c, s_c, and d_c are the number of characters that respectively need to be inserted, substituted, and deleted to reconstruct the transcribed text into the ground-truth.

Similarly, *Word Error Rate* (WER) is calculated as follows:

$$WER = \frac{i_w + s_w + d_w}{n_w} \quad (5)$$

where all the parameters remain the same, except they are counted in words. It is worth noting that WER is generally higher than CER within the same sample, as WER is a stricter evaluation where any character mistake would make a whole word considered as wrong. On the other hand, CER is not as tight as the aforementioned, since the error in character is independent of each other and does not affect any previous or subsequent characters.

The results for both systems are either presented in tables or visualized using boxplots that show the spread of the data by quartiles. In the tables, the results indicated in bold are the best scores obtained by the system according to the type of degradation. We also compute δ as a measure according to degradation type.

$$\delta = F1_{clean\_data} - F1_{noisy\_data} \quad (6)$$

This measure gives the minimum decrease rate between the F1 given using clean data and the F1 given using noisy data for each type of degradation and it represents the perfect system that will give the best F1 for all degradation levels.

The experiments were conducted in the following manner: each noise type is generated with different intensities in order to see the relation between noise intensity and model performance. CER and WER were calculated for each noise level. The experiments are performed under conditions of varying WER and CER: Original text (no OCR, 0% WER, 0% CER); OCR from high-quality text images (∼1% WER, ∼0.5% CER); OCR on degraded text images synthetically produced with DOCCREATOR (2−50% WER, 1−20% CER).

5.3 Hyperparameters

The hyperparameters used for the CNN model for event detection were chosen on the validation set. The window sizes for the convolutional layers used in the experiments are in the set $\{1, 2, 3\}$ to generate feature maps, and 300 feature maps are used for each window size in this set. After each convolutional layer, a *ReLU* activated nonlinear layer is applied with orthogonal weights initialisation. The window size for each trigger candidate that traverses each sentence in the dataset is set to 31 and the dimensionality of the position embeddings is 50 [2]. The batch size is set to 256 and we employed the pre-trained word embeddings of size 300 for *Word2vec* [42][15]. For DAnIEL, the ratio is set to 0.8.

6 Experiments on Daniel-data

For each experiment, we perform two evaluation types:

1. *Event identification*: a document represents an event if the triggers were found, regardless of their types.
2. *Event classification*: a document represents an event if the triggers are correctly found and match with the ground-truth ones.

The data has a total of 4,822 documents, and in order to obtain stable and confident results, we perform the cross-validation resampling method with five experiments for each setting, with around 3,857 documents for training (80%), 482 documents for development (10%), and the rest of 483 documents for testing (10%), stratified by language.

6.1 Experiments with Original Data

Event Identification. For *identifying events* on clean textual data, where the system needs to detect whether a document contains a disease, a location, and the number of victims (a relevant document), one can notice in Table 2, that DANIEL-SYS usually favours the recall instead of precision and tends to suffer from an imbalance between precision and recall, which may be due to the high imbalance of the data. Meanwhile, the CNN-based model is more robust to this characteristic of the

[15] We noticed the fact that the batch size affects the Adam optimizer [60], and thus our choice of 256, which performed the best on the validation set.

Table 3 Evaluation of the CNN-based model and DANIEL-SYS on the original test data for *event classification*.

		Polish	Chinese	Russian	Greek	French	English	All Languages
DANIEL-SYS	P	40.16±6.13	20.00±27.38	24.21±6.90	51.66±3.72	47.59±5.32	50.0±0.0	45.92±3.91
	R	33.71±4.07	20.00±27.38	27.38±9.56	48.74±3.82	54.98±2.15	53.42±4.80	37.12±2.45
	F1	**36.26±2.42**	20.00±27.38	**25.65±8.05**	**50.13±3.54**	50.92±3.56	**51.57±2.20**	**45.89±2.91**
CNN	P	20.00±40.00	28.00±23.15	100.00±0.00	79.56±10.52	64.19±3.29	80.00±40.00	64.75±3.23
	R	1.43±2.85	50.00±44.72	11.67±4.08	12.35±7.05	49.58±11.26	7.69±4.86	26.57±12.49
	F1	2.67±5.33	**34.76±28.87**	20.66±6.46	20.33±10.35	**55.22±7.08**	13.90±8.44	35.91±13.48

dataset. We can note also that DANIEL-SYS seems to detect more relevant documents of lower quality due to the OCR process, giving a higher cost to false positives and favouring in this way recall over precision, for all the cases. The CNN is not able to identify some documents containing events in Russian, Greek, and Polish. The explanation could be that since they are (highly) inflectional languages, all the occurrences of the disease name and location name are not in the same form due to the inflections. Changing automatically the DANIEL-DATA format (document-level annotation) to the more common sentence-level annotation made some events impossible to detect by the CNN. The DANIEL-SYS has a 100% rate of *event identification* for Greek, which could mean that the system managed to detect correctly the six relevant documents (containing a disease-location pair) from the test set.

Event Classification. In the case of *event classification*, where the system needs not only to detect the presence of an event, but also to explicitly detect the disease, location, and the number of victims, we can observe from Table 3, that DANIEL-SYS is still more balanced regarding the precision and recall metrics, being able to have higher F1 on the under-represented languages (Polish, Chinese, Russian, and Greek) than the CNN-based model. DANIEL-SYS leaded to an increase of impressive average gain on Polish, 24.15 average gain on Russian, and most notably, an increase of 146.58% for Greek. Analysing these results, we noticed that, in all the cases, DANIEL-SYS does not detect the number of victims. We assume that this is due to the fact that many of the annotated numbers cannot be found in the text, and thus, the system could not look for repeated substrings at salient zones, e.g. 10000 cannot be detected since the original text has the 10,000 form, or it is spelled "ten thousand". This does not necessarily apply for the CNN-based model, since it can be more robust to word semantics due to the usage of pre-trained word embeddings where 10000 and "ten thousand" can be in the same shared embedding space. We expected higher results for English, due to the availability of resources, however, the drop in percentage from 51.57 to 13.90 is not negligible. The CNN generally performed slightly better than DANIEL-SYS. Improvements in performance were noticed in the case of the CNN model, where this model led to 4.76 average gains on Polish, 16.57 on French, and 7.15 on English. The values of recall for the CNN-based model are in general low. This might be related to the fact that the model is not able to detect some country names due to the fact they are not mentioned in the original text (only a city is mentioned for instance). On the contrary, DANIEL-SYS can detect these locations due to the usage of external resources and article metadata. The small amount of Chinese documents in the testing data were annotated with a disease−location pair, but the locations cannot be found in the text (one of the advantages of DANIEL-SYS of using external resources). DANIEL-SYS is able to detect correctly only the disease names, but the CNN-based model cannot retrieve any of them correctly, even more, the location. Besides this, the small amount of data greatly affects the performance of the CNN-based model. We assume that the CNN-based model performs better for the French documents, due to the larger amount of data, and for the English documents, due to the fact that all the disease−location pairs (in the English documents, no number of victims was annotated) were located in the texts in the exact same form as the annotation, and thus it was easier to correctly detect the event.

6.2 Experiments with Noisy Data

6.2.1 DANIEL-SYS

First, we present the CER (Table 5) and WER (Table 6) values after the noise effects are applied for each language in DANIEL-DATA, as well as for all the documents altogether. For character-based languages (e.g. Chinese), CER is commonly used instead of WER as the measure for OCR, and, thus, we report only the CER [66]. These error values clearly state that *Character Degradation* is the effect that affects the transcription of the documents the most. Moreover, we can easily observe that the highest values for CER, for every type of noise, are obtained for the Chinese documents. We assume that this might be caused by the existence

Table 4 Evaluation of DANIEL-SYS results on the original and the DANIEL-DATA test data for *event identification*. The *Original* results are also presented in Table 2.

		Original	Clean	CharDeg	Bleed	Blur	Phantom	All Effects
Polish	P	80.33±12.26	80.33±12.26	80.00±44.72	80.33±12.26	73.33±18.06	80.33±12.26	80.33±12.26
	R	77.66±13.72	77.66±13.72	29.66±18.94	77.66±13.72	57.00±16.09	77.66±13.72	77.66±13.72
	F1	78.11±8.12	78.11±8.12	42.76±25.82	78.11±8.12	63.76±15.86	78.11±8.12	78.11±8.12
Chinese	P	40.00±54.77	40.00±54.77	40.00±54.77	40.00±54.77	40.00±54.77	40.00±54.77	40.00±54.77
	R	40.00±54.77	40.00±54.77	40.00±54.77	40.00±54.77	40.00±54.77	40.00±54.77	40.00±54.77
	F1	40.00±54.77	40.00±54.77	40.00±54.77	40.00±54.77	40.00±54.77	40.00±54.77	40.00±54.77
Russian	P	76.28±14.61	67.28±12.14	76.28±14.61	76.28±14.61	76.28±14.61	67.28±12.14	81.00±12.33
	R	83.80±10.26	83.80±10.26	83.80±10.26	83.80±10.26	83.80±10.26	83.80±10.26	67.62±10.42
	F1	79.71±12.36	74.49±11.34	79.71±12.36	79.71±12.36	79.71±12.36	74.49±11.34	73.50±10.38
Greek	P	100.0±0.0	84.33±9.39	78.33±12.63	84.33±9.39	100.0±0.0	84.33±9.39	100.0±0.0
	R	100.0±0.0	87.66±11.64	59.00±10.24	87.66±11.64	42.66±7.22	87.66±11.64	54.00±5.47
	F1	100.0±0.0	85.84±9.79	67.23±11.13	85.84±9.79	59.52±7.14	85.84±9.79	70.00±4.56
French	P	56.52±5.61	76.69±2.32	79.34±3.39	76.69±2.32	78.05±2.84	76.69±2.32	81.91±3.35
	R	89.03±2.20	84.81±2.39	77.02±2.06	84.81±2.39	84.81±2.39	84.81±2.39	64.36±3.24
	F1	69.03±4.39	80.52±1.46	78.13±2.22	80.52±1.46	81.26±1.93	80.52±1.46	72.03±2.62
English	P	76.33±5.82	76.33±5.82	60.00±9.13	76.33±5.82	50.00±28.86	76.33±5.82	60.00±9.13
	R	100.0±0.0	100.0±0.0	46.66±7.45	100.0±0.0	36.66±21.73	100.0±0.0	46.66±7.45
	F1	86.47±3.87	86.47±3.87	52.28±7.53	86.47±3.87	42.28±24.77	86.47±3.87	52.28±7.53
All Languages	P	64.08±3.71	76.81±2.17	78.49±3.42	77.68±2.20	77.55±2.47	76.81±2.17	81.65±2.53
	R	89.68±1.94	85.65±1.14	70.36±2.80	85.65±1.14	75.28±2.82	85.65±1.14	63.92±2.53
	F1	74.71±2.78	80.98±1.31	74.19±2.80	81.46±1.27	76.39±2.48	80.98±1.31	71.67±1.78

Table 5 CER degradation values for DANIEL-DATA. *All* is for all effects applied together.

	Clean	CharDeg	Bleed	Blur	Phantom	All
Polish	0.15	5.86	0.19	7.57	0.19	5.51
Chinese	36.89	41.01	38.24	43.97	36.91	46.97
Russian	0.93	16.20	1.45	8.13	1.03	10.91
Greek	3.52	9.04	3.76	13.79	3.54	16.28
French	1.96	8.37	2.13	7.43	2.0	10.90
English	0.35	5.75	0.52	4.74	0.44	7.43
All	2.61	9.55	2.83	8.76	2.65	11.07

Table 6 WER degradation values for DANIEL-DATA. *All* is for all effects applied together.

	Clean	CharDeg	Bleed	Blur	Phantom	All
Polish	0.74	20.66	1.17	13.23	1.17	20.70
Chinese	–	–	–	–	–	–
Russian	1.63	28.46	6.61	14.94	2.73	29.72
Greek	15.86	41.36	17.39	54.02	15.93	54.76
French	3.33	23.56	4.89	16.31	3.76	26.07
English	0.66	24.78	2.14	14.72	1.66	20.99
All	4.23	26.23	5.93	19.05	4.71	27.36

of the enormous number of characters in the alphabet that, by adding any type of noise, can change drastically the recognition of a character (moreover, in Chinese, one single character can often be a word). While *Character Degradation* noise and *Blur* effect have more impact on the CER and WER of DANIEL-DATA, *Phantom Character* and *Bleed Through* have little to no clear visibility. We think that this is because these types did not generate enough distortion to the images[16].

[16] This is also observed in Figure 7.

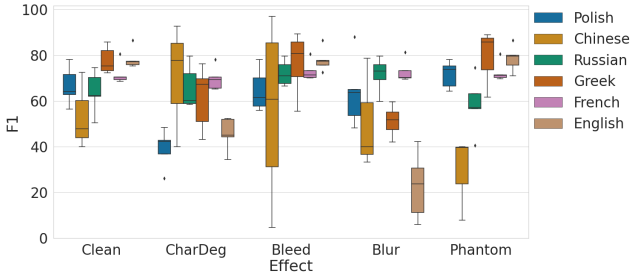

Fig. 8 The distribution of DANIEL-SYS F1 scores on the noisy test DANIEL-DATA for *event identification*.

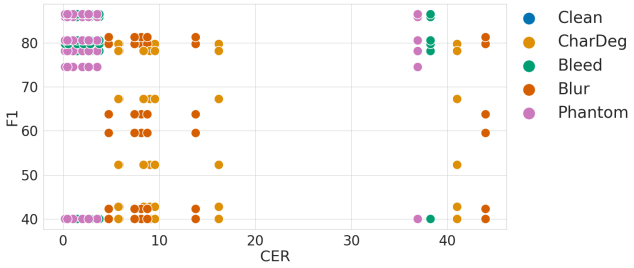

Fig. 9 The distribution of DANIEL-SYS F1 scores in regards to the CER values on the noisy test DANIEL-DATA for *event identification*.

Event Identification. For *event identification* from noisy documents over DANIEL-DATA with DANIEL-SYS, we present the detailed results in Table 4, the distribution of F1 scores in Figure 8, and this distribution in regards to the CER values in Figure 9. In Figure 8, we notice that the *Character Degradation* effect, *Blur*, and most of all, all the effects mixed together, have indeed

Table 7 Evaluation of DANIEL-SYS results on the original and the DANIEL-DATA test data for *event classification*.

		Original	Clean	CharDeg	Bleed	Blur	Phantom	All Effects
Polish	P	40.16±6.13	40.16±6.13	40.00±22.36	40.16±6.13	36.66±9.03	40.16±6.13	40.16±6.13
	R	33.71±4.07	33.71±4.07	12.41±7.66	33.71±4.07	24.49±5.32	33.71±4.07	33.71±4.07
	F1	36.26±2.42	36.26±2.42	18.79±11.17	36.26±2.42	29.27±6.32	36.26±2.42	36.26±2.42
Chinese	P	20.00±27.38	20.00±27.38	20.00±27.38	20.00±27.38	20.00±27.38	20.00±27.38	20.00±27.38
	R	20.00±27.38	20.00±27.38	20.00±27.38	20.00±27.38	20.00±27.38	20.00±27.38	20.00±27.38
	F1	20.00±27.38	20.00±27.38	20.00±27.38	20.00±27.38	20.00±27.38	20.00±27.38	20.00±27.38
Russian	P	24.21±6.90	26.96±4.54	30.46±4.57	30.46±4.57	17.03±9.80	26.96±4.54	31.00±12.33
	R	27.38±9.56	33.80±5.21	33.80±5.21	33.80±5.21	19.28±10.89	33.80±5.21	25.71±8.98
	F1	25.65±8.05	29.93±4.56	31.98±4.59	31.98±4.59	18.06±10.25	29.93±4.56	28.03±10.31
Greek	P	51.66±3.72	37.66±10.04	25.00±5.89	37.66±10.04	25.0±0.0	37.66±10.04	50.0±0.0
	R	48.74±3.82	36.67±8.96	17.65±3.81	36.67±8.96	10.07±1.84	36.67±8.96	25.51±3.17
	F1	50.13±3.54	37.10±9.28	20.65±4.52	37.10±9.28	14.28±1.88	37.10±9.28	33.69±2.74
French	P	47.59±5.32	65.20±2.28	70.50±3.32	65.20±2.28	68.66±2.45	65.20±2.28	72.51±4.72
	R	54.98±2.15	52.91±1.00	50.22±1.40	52.91±1.00	54.76±1.54	52.91±1.00	41.81±2.95
	F1	50.92±3.56	58.38±0.52	58.63±1.63	58.38±0.52	60.89±1.12	58.38±0.52	53.01±3.32
English	P	50.0±0.0	50.0±0.0	29.99±4.56	50.0±0.0	24.99±14.43	50.0±0.0	29.99±4.56
	R	53.42±4.80	53.42±4.80	18.85±2.55	53.42±4.80	14.85±8.66	53.42±4.80	18.85±2.55
	F1	51.57±2.20	51.57±2.20	23.08±2.99	51.57±2.20	18.63±10.82	51.57±2.20	23.08±2.99
All Languages	P	45.92±3.91	55.32±2.63	58.63±4.30	55.96±3.03	56.68±3.67	55.32±2.63	60.18±4.49
	R	50.64±2.54	48.62±0.77	41.41±2.32	48.62±0.77	43.35±2.13	48.62±0.77	37.12±2.45
	F1	48.11±2.94	51.74±1.44	48.52±2.86	52.02±1.64	49.11±2.59	51.74±1.44	45.89±2.91

an impact or effect over the performance of DANIEL-SYS, with high variability for Chinese and Polish. Meanwhile, *Phantom Degradation* and *Bleed through* had very little to no impact on the quality of detection with DANIEL-SYS. Figure 9 clearly shows that the decrease of performance is direct proportional with the increase in CER values with the F1 values for English and Chinese are clearly divided from the other languages. We need to emphasize that we observed that the F1 scores for Chinese are not reliable due to the limited amount of annotated events present in our the `Daniel-data`[17]. The cause of the decrease in performance of DANIEL-SYS is that, in order to detect events, the system looks for repeated substrings at salient zones. In the case of many incorrectly recognised words during the OCR process, there may be no repetition anymore, implying that the event will not be detected. However, since DANIEL-SYS only needs two occurrences of its clues (substring of a disease name and substring of a location), it is assumed to be robust to the loss of many repetitions, as long as two repetitions remain in salient zones.

Event Classification. For *classifying events* from noisy documents over DANIEL-DATA with DANIEL-SYS, in Table 7, all the scores drop considerably. Figure 10 reveals the same expected observation. We notice that for Polish and Russian, the F1 values for each of the noise effect, and as well for all mixed together, decrease in comparison with the F1 for original documents. These two collections contain documents in languages with diacritics (e.g., ó, ź, ę, ł, etc., for Polish, $\overline{\alpha}$, $\acute{\epsilon}$, ç, $\acute{\omega}$ for Greek). These characters are typically among the ones with more extraction errors, the insertion of spurious symbols. Figure 11 shows the same tendency as Figure 9, where generally, *Character Degradation* and *Blur* have larger variations in CER and F1 than *Phantom Degradation*. Since *Character Degradation* adds synthetic ink dots to random letters, these are easily confused with

[17] We designate this a limitation in our setup and we detail it in Section 8.

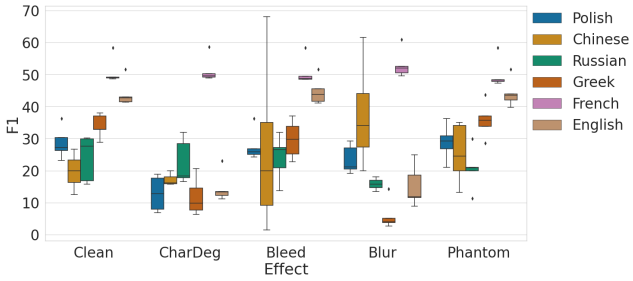

Fig. 10 The distribution of DANIEL-SYS F1 scores on the noisy test DANIEL-DATA for *event classification*.

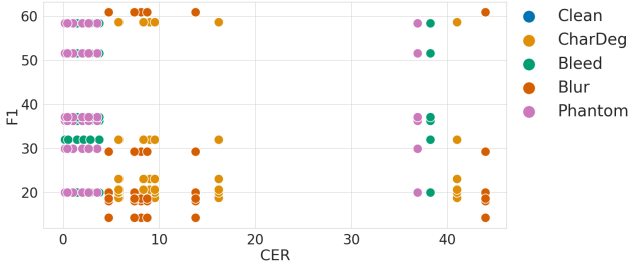

Fig. 11 The distribution of DANIEL-SYS F1 scores in regards to the CER values on the noisy test DANIEL-DATA for *event classification*.

Table 8 Evaluation results of the CNN-based model on the original and the noisy DANIEL-DATA test data for *event identification*.

		Original	Clean	CharDeg	Bleed	Blur	Phantom	All Effects
Polish	P	84.60±18.45	68.68±23.05	69.94±24.49	62.43±19.36	75.97±21.74	74.38±20.67	69.95±24.49
	R	54.64±3.11	52.86±4.86	51.20±1.60	51.07±1.42	55.00±5.69	53.93±4.57	51.36±1.81
	F1	57.93±5.21	54.55±7.55	52.22±2.97	51.76±2.30	57.84±8.64	56.37±6.91	52.51±3.35
Chinese	P	63.00±10.77	52.85±5.71	61.33±9.33	59.16±10.00	61.00±2.90	63.05±6.85	50.00±0.00
	R	80.00±24.49	60.00±19.99	79.99±24.49	79.98±24.48	99.99±0.01	89.99±19.99	50.00±0.00
	F1	68.09±14.87	54.44±8.88	66.42±13.47	63.20±13.59	67.85±3.68	69.60±10.07	50.00±0.00
Russian	P	91.64±10.53	84.63±18.45	94.64±6.86	72.97±22.71	79.65±16.94	86.64±8.07	49.97±0.00
	R	68.33±9.34	56.92±5.10	61.00±4.89	59.22±8.95	61.43±9.68	68.45±7.45	50.00±0.00
	F1	72.27±5.53	61.02±7.34	66.88±5.63	60.99±9.55	65.38±11.01	72.53±5.98	49.98±0.00
Greek	P	80.77±15.77	62.34±12.28	86.63±19.43	67.16±12.09	49.97±0.00	67.45±9.28	49.97±0.00
	R	59.99±7.28	54.34±3.63	55.00±3.18	58.07±4.66	50.00±0.00	56.51±4.55	50.00±0.00
	F1	64.26±8.88	56.31±5.47	58.64±5.34	60.32±6.52	49.98±0.00	59.20±6.13	49.98±0.00
French	P	82.88±0.70	80.70±3.86	77.38±4.44	81.30±4.19	81.83±4.06	81.02±1.12	84.95±7.59
	R	75.83±6.74	75.16±10.84	74.84±14.43	75.93±10.89	82.77±10.83	80.68±7.81	62.12±11.35
	F1	78.37±4.47	76.05±5.39	73.52±11.22	76.41±5.93	80.39±5.46	80.19±3.71	64.57±11.63
English	P	97.88±2.63	88.19±19.42	99.97±0.00	89.95±20.00	75.31±5.81	99.05±1.81	69.97±24.49
	R	61.92±5.62	57.31±9.90	54.38±1.53	55.00±4.48	57.85±8.57	58.08±6.00	51.18±1.44
	F1	68.29±7.04	60.09±11.53	57.96±2.56	58.50±6.92	60.75±9.20	62.89±7.86	52.20±2.72
All Languages	P	82.98±0.74	80.92±3.70	79.80±2.36	80.49±4.50	81.40±3.43	80.77±1.16	85.28±7.34
	R	71.16±5.44	70.09±9.15	70.83±12.10	70.89±9.07	77.71±9.32	75.06±6.48	59.65±8.93
	F1	75.34±4.19	72.85±5.19	71.88±10.21	72.93±5.80	77.84±5.55	77.04±3.85	62.60±10.16

acute accents (ć, ń, ó, ś, ź), the overdot (ż), the tail (ę) or the stroke (ł). The results showed that character error rates starting at around 5% can cause a significant impact in the Polish and Greek system configurations and that stemming makes systems more robust to coping with errors. For Polish, the highest CER reached 7.57%, while for Greek, it started from 9.04% and increased to 16.28%. We also note that even that Russian also is characterised by the presence of diacritics (ë, й, etc.), the impact on the performance is considerably reduced. We assume that one of the reasons is that, in comparison with the other two aforementioned languages, in Russian, there are only two diacritical signs, "˘" which is found only above the letter й, and "¨" which is put only above the letter ë. Thus, besides the fact that because the ë is the least used letter in the Russian alphabet, these accents are generally rarer, and thus, they have a marginal OCR impact in regard to diacritics. The performance scores fluctuate more than in the case for *event identification*, depending on the noise type (as presented in Figure 10), having from fairly dispersed boxplots to largely dispersed ones (e.g. *Blur*). Similar observations as for *event identification* can also be noticed when the DANIEL-SYS has to specifically detect the events, regarding the most impactful noise effects, *Character Degradation* and *Blur*, and the least impactful, *Bleed Through* and *Phantom*). Figure 10 shows this variation, with the highest F1 scores for detecting events in original documents, followed by the three least impacted articles (*Clean*, *Bleed Through*, and *Phantom Character*). For the *Blur* documents, the median is the lowest, which indicates that the level of this

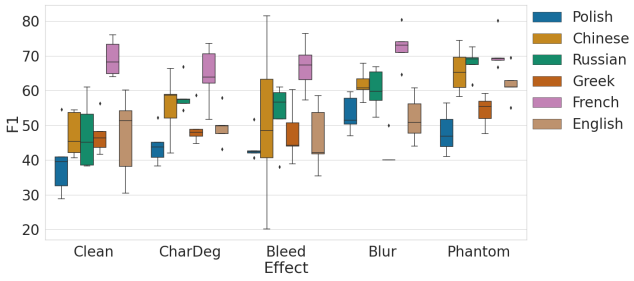

Fig. 12 The distribution of the CNN-based model results on the noisy test DANIEL-DATA for *event identification*.

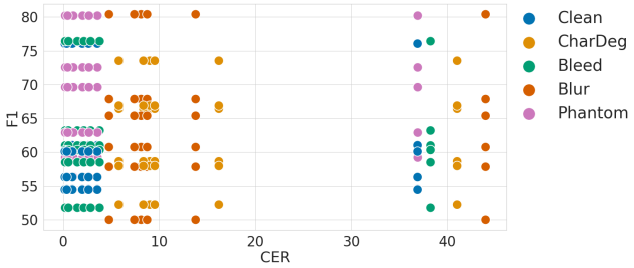

Fig. 13 The distribution of CNN F1 scores in regards to the CER values on the noisy test DANIEL-DATA for *event identification*.

effect is the most influential between all the noise effects, by introducing deletion errors and thus, missing letters impact the performance of DANIEL-SYS having as a main consequence the loss of many repetitions, making it more difficult to detect the events in document.

Table 9 Evaluation results of the CNN-based model on the original and the noisy DANIEL-DATA test data for *event classification*.

		Original	Clean	CharDeg	Bleed	Blur	Phantom	All Effects
Polish	P	20.00±40.00	20.00±40.00	0.00±0.00	53.33±45.21	67.49±36.61	20.00±40.00	20.00±40.00
	R	1.43±2.85	2.14±4.28	0.00±0.00	4.29±4.16	16.43±15.08	0.71±1.42	1.82±3.63
	F1	2.67±5.33	3.87±7.74	0.00±0.00	7.83±7.51	23.23±18.95	1.38±2.75	3.33±6.66
Chinese	P	28.00±23.15	23.33±12.24	16.67±21.08	32.67±17.17	16.33±8.58	10.71±13.17	0.00±0.00
	R	50.00±44.72	40.00±20.00	20.00±24.49	70.00±40.00	80.00±40.00	30.00±40.00	0.00±0.00
	F1	34.76±28.87	29.33±14.96	18.00±22.27	44.19±23.71	27.05±14.00	15.56±19.37	0.00±0.00
Russian	P	100.00±0.00	70.00±40.00	76.00±38.78	51.00±42.47	55.33±32.35	80.33±16.74	81.67±15.27
	R	11.67±4.08	20.00±15.83	18.00±13.26	18.46±17.94	31.43±16.65	21.54±7.53	26.00±18.54
	F1	20.66±6.46	27.46±17.61	27.64±17.76	25.65±23.07	39.23±20.66	32.50±9.56	34.63±16.95
Greek	P	79.56±10.52	32.24±19.88	70.00±40.00	31.33±17.61	20.00±40.00	11.43±22.85	45.00±45.82
	R	12.35±7.05	8.70±5.49	8.75±6.37	15.24±9.23	4.62±9.23	3.48±6.95	4.62±3.76
	F1	20.33±10.35	13.52±8.39	15.34±10.74	19.91±11.66	7.50±15.00	5.33±10.66	8.07±6.65
French	P	64.19±3.29	47.95±24.31	60.68±5.64	64.32±6.09	60.78±9.75	63.28±6.65	68.32±6.35
	R	49.58±11.26	49.85±29.35	50.00±8.33	57.72±19.02	66.62±24.86	62.00±17.60	32.21±15.77
	F1	55.22±7.08	47.55±24.98	54.46±5.83	58.00±10.17	59.04±12.74	60.50±7.46	40.68±15.36
English	P	80.00±40.00	80.00±40.00	76.00±38.78	96.67±6.66	62.78±16.25	100.00±0.00	20.00±40.00
	R	7.69±4.86	9.23±5.21	8.75±8.47	9.23±5.75	24.29±21.47	13.08±6.70	2.35±4.70
	F1	13.90±8.44	16.47±9.11	14.68±12.56	16.21±8.99	32.32±24.73	22.52±10.22	4.21±8.42
All Languages	P	64.75±3.23	47.87±24.24	61.19±5.57	63.96±6.17	60.55±8.97	63.12±6.24	68.11±6.25
	R	38.19±8.83	39.50±22.82	40.39±7.25	46.30±15.55	57.55±22.35	48.40±13.74	26.57±12.49
	F1	47.46±7.05	42.26±22.18	48.31±5.75	51.16±10.36	54.82±12.72	53.04±7.79	35.91±13.48

6.2.2 CNN-based Model

Event Identification. For *event identification* from noisy documents with the CNN-based model over DANIEL-DATA, in Table 8, all collections for each language are impacted, having the highest performance score for the original data. We excluded the results for the Chinese documents, since, for both evaluation types, the values were equal to zero. The decrease in precision and recall is similar to DANIEL-SYS, the impact on the scores being higher for the *Character Degradation*, *Blur*, and all mixed together, also. There are also cases where F1 was set to zero, with no event identified, i.e. *Character Degradation* for Polish, and *Blur* for Greek and English.

Figure 12 shows that the F1 scores vary considerably for each language. Nonetheless, we can observe that the median is consistent for *Clean*, *Character Degradation*, *Bleed Through*, and *Phantom Character*, but not for *Blur*, and all effects mixed together. This indicates that, while each noise applied to the documents decreases the F1 performance, when applied together, the effectiveness of the model drops significantly. The distribution of F1 scores in regards to the CER values from Figure 13 show also that *Blur* and *Character Degradation* tend to have higher CER values with various performance scores. One drawback of this model is that it is based on embeddings at the word level, which can degrade the performance in the case of many modified words in the test set during the OCR process. While this can contribute to the lowering of the scores, we remind that, since the dataset is multilingual, we randomly generated the word embeddings, thus, the misspelled words will also have a meaningful representation. However, due to the OCR errors, the size of the vocabulary also increases and the semantics of misspelled words can notably change due to variations of context.

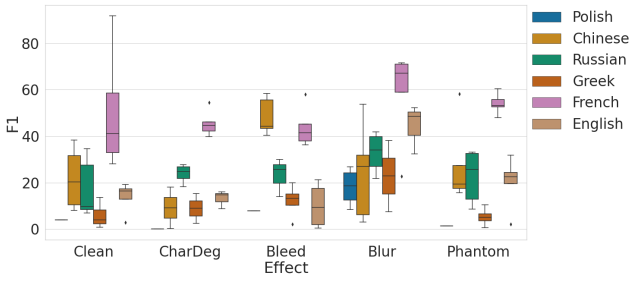

Fig. 14 The distribution of the CNN-based model results on the noisy test DANIEL-DATA for *event classification*.

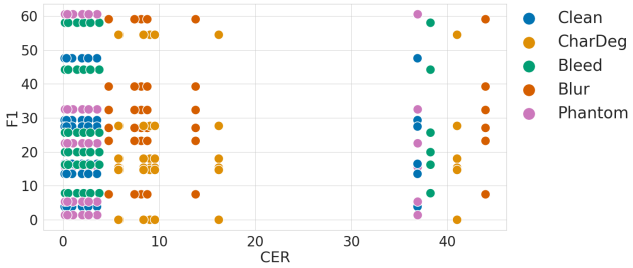

Fig. 15 The distribution of CNN F1 scores in regards to the CER values on the noisy test DANIEL-DATA for *event classification*.

Event Classification. For *event classification* from noisy documents over DANIEL-DATA with the CNN-

based model, in Table 9, we observe that, besides the fact that the F1 scores decreased due to the difficulty introduced by the complexity of the task, the median is slightly similar for the clean, *Character Degradation*, *Blur*, and *Phantom Character* collections, while the highest, as expected, is on the original dataset. Figure 14 shows that the distribution of F1 scores is rather similar to the ones for *event identification*, thus, maintaining a slightly comparable level of variability, depending on the noise effect. All effects mixed together affect greatly the distribution of the F1 scores due to the fact the Polish, Russian, and English collections were substantially corrupted and cause the scores to drop to zero.

6.2.3 Discussion

Studying the degree of variability of F1-scores for all the effects mixed together for event identification and classification, we notice the CNN-based model is more sensitive to the added effects, as shown in Figure 16. We conclude that using representations at the word level in the CNN-based model indeed hurts the performance of the model when evaluated on the text transcribed from degraded images.

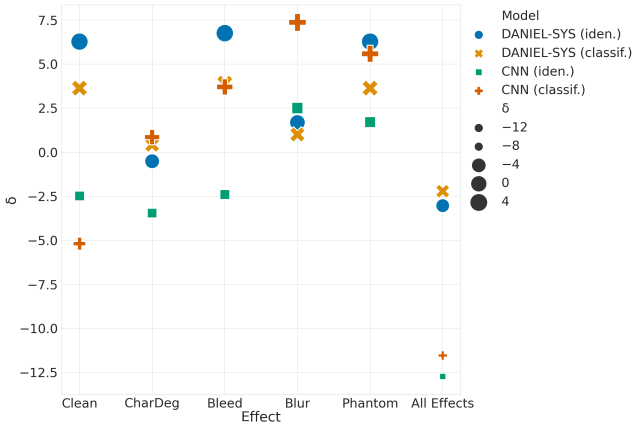

Fig. 16 δ degradation values according to OCR error rates for both models by each noise effect, and for all effects, as well, for event identification (iden.) and classification (classif.).

Figure 16 shows that δ is mostly positive and with higher values for the DANIEL-SYS model, for all the languages. For the CNN-based model, some values are marginally positive, more exactly, for *Character Degradation*, for both *event identification* and *event classification*. For *All effects*, the CNN model has very low δ values (right lower corner in Figure 16, which proves that this type of model can highly be influenced by the digitisation process.

For both *event identification* and *event classification*, DANIEL-SYS is more prone to produce a positive δ. This means that the F1 scores decrease after the noise is applied, while for the negative values, the precision and recall scores actually increase. This brings us to an interesting observation that the scores can increase, resulting in a higher F1, despite the higher noise effect applied, for both *event identification* and *event classification* with DANIEL-SYS.

The main reason for this unexpected result is that with a greater level of noise some false positives disappear. Documents, which were previously wrongly classified due to being too ambiguous to the system (for instance, as mentioned before, documents related to vaccination campaigns are usually annotated as irrelevant in DANIEL-DATA), were given much more distinction due to the noise, thus making them look less like relevant samples to the system. This may seem counterintuitive but noise can improve classification results, see for instance [34] for a study on the same dataset of the influence of boilerplate removal on results.

Regarding all the results aforementioned, for the DANIEL-SYS, and the CNN-based model, computing the number of affected event words (disease, location, number of patients), we also notice that a very small number of them have been modified by the OCR process, only 1.98% for all the languages together, for all the effects mixed together, not far from the 1.63% that were affected by the OCR on clean data. This is due to the fact that DANIEL-DATA is highly imbalanced (only 10.14% of a total of 4,822 documents do contain events), and it brings us to the conclusion that the event detection task is not considerably impacted by the degradation of the image documents.

Both methods cover two research lines for detecting events, more specifically, an unsupervised method that does not rely on labels or feature engineering [33], and supervised learning that, while it needs annotated data, it takes advantage of word representations that are learned on large corpora [2]. However, we would expect that methods that are based on pre-trained language models could improve the general performance, but the fact that they still rely on a pre-set vocabulary would generate the same variation of the results [5][18].

[18] While Bidirectional Encoder Representations from Transformers (BERT) had a major impact in the NLP community, its ability to handle noisy inputs is still an open question [63] or at least requires the addition of complementary methods [44,52].

Table 10 Evaluation of the CNN-based model on the original ACE 2005 test data for *event identification* and *classification*.

CNN [48] (reported)			Our CNN (replicated)		
P	R	F1	P	R	F1
Event identification					
-	-	-	71.64±1.35	77.58±0.77	74.02±0.60
Event classification					
71.90	63.80	67.60	68.88 ±0.69	58.45 ±1.56	63.18 ±0.91

Table 11 Evaluation results on the noisy ACE 2005 test data for *event identification*. CharDeg = character degradation, Bleed = Bleed through, All = CharDeg + Bleed + Phantom + Blur.

	Original	Clean	CharDeg	Bleed	Blur	Phantom	All
P	71.64±1.35	70.09±0.54	57.38±0.62	67.07±0.41	56.6±0.92	64.99±0.81	54.44±1.40
R	77.58±0.77	76.99±0.99	77.33±0.52	77.55±0.85	73.25±0.31	79.28±0.77	74.45±0.52
F1	74.02±0.60	72.92±0.45	60.62±0.85	70.89±0.16	59.08±1.25	69.43±0.69	54.96±2.34

Table 12 Evaluation results on the noisy test data for *event classification*. CharDeg = character degradation, Bleed = Bleed through, All = CharDeg + Bleed + Phantom + Blur.

	Original	Clean	CharDeg	Bleed	Blur	Phantom	All
P	68.82±0.83	68.62±1.23	47.63±1.09	57.75±1.05	67.55±1.30	59.05±1.52	48.02±0.79
R	66.13±1.24	65.51±0.78	50.54±0.92	64.37±0.87	53.77±1.19	64.94±1.34	35.48±0.66
F1	67.40±0.51	66.97±0.14	48.97±0.44	60.82±0.20	59.80±0.59	61.72±0.30	40.77±0.36

7 Experiments on ACE 2005 Dataset

We perform the following evaluation from the ACE 2005 evaluation [48,47]:

1. *Event identification*: a sentence represents an event if the triggers were found, regardless of their types;
2. *Event classification*: a sentence represents an event if the triggers are correctly found and match with the ground-truth ones.

Since for ACE 2005 the data split is well-established in previous research [13,12,2], we do not perform cross-validation resampling but, we run five experiments with different random seeds in order to obtain stable results.

7.1 Experiments with Original Data

For the experiments with clean ACE 2005 data, we replicated the model presented in [48,2] and the results for *event classification* are presented in Table 10. We can notice that our performance is close to the one reported in previous research. We also add the results for *event identification*.

7.2 Experiments with Noisy Data

First, Table 13 presents the error rates for each noise effect, along with the percentage of event triggers that were modified due to the OCR process. Next, we present the results for *event identification* and *event classification* for the CNN-based model on the ACE 2005 collections: Original (original dataset), Clean (the transcribed documents after applying the OCR), and the documents after each of the noise effects (*Character Degradation*, *Bleed Through*, *Blur*, and *Phantom Character*, and *All mixed together*). From the results for *event identification* in Table 11 and *event classification* in Table 12, we notice the same tendency, the fact that the impact on the evaluation scores are higher for the *Character Degradation*, *Blur*, and all mixed together, than for *Bleed Through* and *Phantom Character*, results that coincide to our observations on Daniel-data. Figure 17 presents the variability of F1 scores with regard to the applied effect, and it is visible that *All effects* applied together can cause changeability in scores. However, for *classification*, this variability is less visible in Figure 18.

Table 13 CER and WER evaluation values for the noisy ACE 2005 test data. The percentage of affected triggers (%) is also presented. CharDeg = character degradation, Bleed = Bleed through, Orig. = Original, All = CharDeg + Bleed + Phantom + Blur.

	Orig.	Clean	CharDeg	Bleed	Blur	Phantom	All
CER	-	0.83	4.10	1.34	7.28	0.95	14.81
WER	-	1.13	17.96	5.61	18.49	2.50	35.93
Affected triggers	-	0.94	19.05	2.11	19.05	0.94	41.17

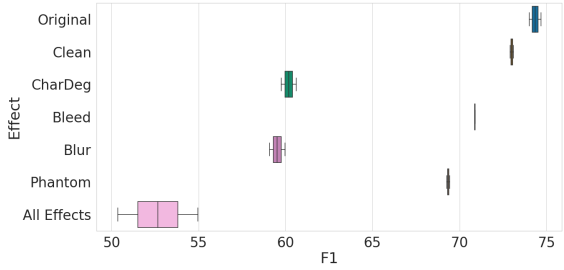

Fig. 17 The distribution of the CNN-based model F1 scores on the noisy test ACE 2005 for *event identification*.

We recall that one drawback of this model is that it is based on a pre-defined set of word embeddings, which can degrade the performance in the case of many wrongly detected words in the OCR process. The results, however, are consistent with the drop in the quality of the documents, and thus, for the two highest values of CER, 4.10 for *Character Degradation* and 14.81 for all the noise effects together, the lowest F1 values were obtained, 48.97, and 40.77 respectively.

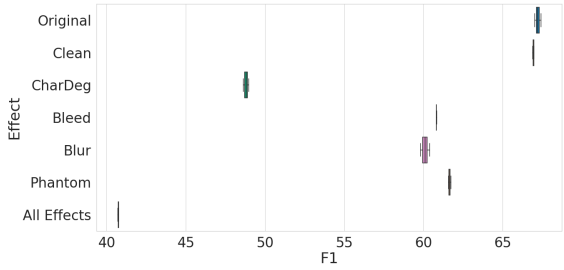

Fig. 18 The distribution of the CNN-based model F1 scores on the noisy test ACE 2005 for *event classification*.

After applying the synthetic noise effects to the images and the digitisation process, we analysed the transcribed event triggers, in order to assess the level of their degradation. We observed that the number of event triggers that were affected by the OCR process when all the noise levels were applied is 41.17% out of all event triggers, and thus, this justifies the large drop of around 27 percentage points. Also, while 19.05% of the event triggers were affected in two cases, *Character Degradation* and *Blur*, the CER error rates (4.10 and 7.28, respectively) and the F1 values differ (48.97% and 59.50% respectively). An explanation is that the precision of the results in the case of the *Blur* is considerably higher than in the case of *Character Degradation*, which would mean that even though both models managed to retrieve a similar amount of event triggers (a recall of 50.54% and a recall of 53.77%), the CNN-based models were able to better detect the correct event type even when the words were affected by the *Character Degradation* noise.

8 Discussion

Datasets. An important observation that should be noted is regarding the imbalance of both datasets, a common characteristic of event-annotated corpora. When comparing the digitisation impact onto the Daniel-data and ACE 2005, we can safely conclude that the misproportion of labeled events is an important factor in assessing the level of influence that the OCR process can have on the event detection task.

Daniel-data is highly imbalanced and the variability in results was lower than in the case of the ACE 2005. Consequently, the probability that the few words annotated as events were affected was quite low. ACE 2005 has a much higher number of events and event types in almost every document. 92.32% of its documents are relevant by containing one or multiple events of different types, while in the Daniel-data, only 10.14% have events of a single type (epidemic). Therefore, concurrently, we can also declare that the richness of the annotation styles held a high impact on the performance of both models.

Between the noise effects that were applied on the documents, *Character Degradation* and *Blur* had a significant effect on the performance of both models when compared to *Bleed Through* and *Phantom Character* that are slightly visible, henceforth, it is safe to say that this conclusion could translate to other event-based datasets, regardless of the event definition and types. However, when it comes to the multilinguality of the datasets, the impact of the OCR can also be assessed. Daniel-data covers news articles from several, diverse language families; Germanic: English (en), Hellenic: Greek (el), Romance: French (fr), Slavic: Russian and Polish, and Chinese that descends from the Sino-Tibetan family. From these, Greek, Russian, and Polish, not only that they make use of a high degree of inflection, but they also contain terms that have letters with diacritical marks. These two factors contributed to the degree of degradation after the digitisation by increasing the number of character errors, either by addition, substitution or deletion.

Models. Overall, the CNN-based model is more impacted by the effects of the noise added to the images than the Daniel-sys. We believe this is due to the more robust string-level representation used by the Daniel-sys, compared to the word-level representation of other approaches. The CNN-based model, even though it automatically generates randomly initialised embeddings

for the misspelled or affected words, their heterogeneity makes them lose meaning during the training process, and thus the model becomes less able at identifying the correct event type. The fact that the noise has less impact on the DANIEL-SYS results can also be explained by the fact that the model uses external resources in order to predict the presence of an event. One disadvantage of this model might be its exclusive applicability to epidemic events, and the amount of effort needed in order to adapt it to other domains (e.g. Wikipedia seeds for different domains need to be provided). An advantage that is common to both models is language independence. If we compare the gap between the scores for *event identification* and *event classification*, both models maintain a slightly similar distribution of scores for the noise effects, and, as expected, the performance decreased proportionally with the difficulty of extracting the events of specific types.

Limitations. While our experimental setup covers the variability of event detection annotation styles, different dataset sizes, languages, and types of noise, the results only apply for a specific setting: Arial font, Tesseract OCR tool v4.0[19], and the types of noise with a single set of effect parameters. Moreover, the lack of Chinese annotated events in the case of DANIEL-DATA could lack generalization. Nonetheless, the imbalance still remains an important factor to consider, along with the complexity of the models.

9 Conclusions

We conclude that, in general, event detection is indeed prone to errors induced by an imperfect OCR, depending on the level of data imbalance, the annotation style, language characteristics, text representation, and of course, the type of noise applied to the documents.

The noisy documents that were produced for the two event detection datasets have been aligned with their corresponding ground-truth in order to further experiment with other event detection systems through noisy data and to observe the evolution of their performance. The DANIEL-DATA dataset[20] is made publicly available to the community[21].

This type of study and such resources that combine digitised data aligned with their clean version is very useful for the following reasons. First, they can leverage the recent advances in applying event detection over collections of digitised and/or historical documents [61]. Second, they can be utilised to estimate the impact of the OCR process on this task and to reduce the human expertise and manual labor-intensive work for hand-validating transcribed documents. This can also lead to recommendations, for instance, on what application can reasonably be applied over a document collection given its OCR quality. Third, they make it easier for the community to address the important challenge of the lack of ground-truth data with high variability (as shown to be problematic by numerous works on event detection in historical documents [30, 9, 10, 56]).

[19] `https://github.com/tesseract-ocr/tesseract`

[20] The ACE 2005 dataset is available under a paid license and thus, we cannot make it available.

[21] `https://zenodo.org/record/3709617`

Conflict of interest

The authors declare that they have no conflict of interest.

Author contributions

Emanuela Boros: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data Curation, Writing. **Nhu Khoa Nguyen:** Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing. **Gaël Lejeune:** Conceptualization, Methodology, Supervision, Writing - review & editing. **Antoine Doucet:** Funding acquisition, Conceptualization, Methodology, Project administration, Validation, Supervision, Writing - review & editing.

Availability of code, data and material

Our code is freely available at `https://github.com/NewsEye/event-detection`.

Acknowledgements This work has been supported by the European Union's Horizon 2020 research and innovation program under grants 770299 (NewsEye) and 825153 (EMBEDDIA), and by the ANNA and Termitrad projects funded by the Nouvelle-Aquitaine Region.

References

1. 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)
2. Boros, E.: Neural methods for event extraction. Ph.D. thesis, Université Paris Sud (2018)
3. Boros, E., Besançon, R., Ferret, O., Grau, B.: The importance of character-level information in an event detection model. In: International Conference on Applications of

Natural Language to Information Systems, pp. 119–131. Springer (2021)
4. Boroş, E., Besançon, R., Ferret, O., Grau, B.: Intérêt des modèles de caractères pour la détection d'événements (the interest of character-level models for event detection). In: Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1: conférence principale, pp. 179–188 (2021)
5. Boros, E., Hamdi, A., Linhares Pontes, E., 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. Association for Computational Linguistics, Online (2020). DOI 10.18653/v1/2020.conll-1.35. URL https://www.aclweb.org/anthology/2020.conll-1.35
6. Boros, E., Linhares Pontes, E., Cabrera-Diego, L.A., Hamdi, A., Moreno, J.G., Sidère, N., Doucet, A.: Robust Named Entity Recognition and Linking on Historical Multilingual Documents. In: L. Cappellato, C. Eickhoff, N. Ferro, A. Névéol (eds.) CLEF 2020 Working Notes. Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum. CEUR-WS (2020)
7. Boros, E., Moreno, J., Doucet, A.: Event detection with entity markers. In: European Conference on Information Retrieval, pp. 233–240. Springer (2021)
8. Boroş, E., Romero, V., Maarand, M., Zenklová, K., Křečková, J., Vidal, E., Stutzmann, D., Kermorvant, C.: 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)
9. Boschee, E., Natarajan, P., Weischedel, R.: Automatic extraction of events from open source text for predictive forecasting. In: Handbook of Computational Approaches to Counterterrorism, pp. 51–67. Springer (2013)
10. Boschetti, F., Cimino, A., Dell'Orletta, F., Lebani, G., Passaro, L., Picchi, P., Venturi, G., Montemagni, S., Lenci, A.: 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)
11. Bronstein, O., Dagan, I., Li, Q., Ji, H., Frank, A.: Seed-based event trigger labeling: How far can event descriptions get us? In: ACL (2), pp. 372–376 (2015)
12. Chen, C., Ng, V.I.: Joint modeling for chinese event extraction with rich linguistic features. In: In COLING. Citeseer (2012)
13. 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)
14. Collier, N.: Towards cross-lingual alerting for bursty epidemic events. Journal of Biomedical Semantics **2**(5), S10 (2011)
15. Collier, N., Doan, S., Kawazoe, A., Goodwin, R.M., Conway, M., Tateno, Y., Ngo, Q.H., Dien, D., Kawtrakul, A., Takeuchi, K., et al.: Biocaster: detecting public health rumors with a web-based text mining system. Bioinformatics **24**(24), 2940–2941 (2008)
16. Cybulska, A., Vossen, P.: Event models for historical perspectives: Determining relations between high and low level events in text, based on the classification of time, location and participants. In: LREC (2010)
17. 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)
18. Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., Weischedel, R.: The automatic content extraction (ace) program–tasks, data, and evaluation. In: Proceedings of LREC, vol. 4, pp. 837–840. Citeseer (2004)
19. Du, X., Cardie, C.: Event extraction by answering (almost) natural questions. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 671–683. Association for Computational Linguistics, Online (2020). DOI 10.18653/v1/2020.emnlp-main.49. URL https://aclanthology.org/2020.emnlp-main.49
20. Duan, S., He, R., Zhao, W.: Exploiting document level information to improve event detection via recurrent neural networks. In: Eighth International Joint Conference on Natural Language Processing (IJCNLP 2017), pp. 352–361. Asian Federation of Natural Language Processing (2017)
21. 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 (Volume 2: Short Papers), vol. 2, pp. 66–71 (2016)
22. Filatova, E., Hatzivassiloglou, V.: Event-based extractive summarization (2004)
23. Giguet, E., Lucas, N.: La détection automatique des citations et des locuteurs dans les textes informatifs. Le discours rapporté dans tous ses états: Question de frontières pp. 410–418 (2004)
24. Grishman, R., Sundheim, B.: Message understanding conference-6: A brief history. In: COLING 1996, pp. 466–471 (1996)
25. Hamborg, F., Lachnit, S., Schubotz, M., Hepp, T., Gipp, B.: Giveme5w: Main event retrieval from news articles by extraction of the five journalistic w questions. In: Transforming Digital Worlds, pp. 356–366. Springer International Publishing, Cham (2018). DOI 10.1007/978-3-319-78105-1_39
26. 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)
27. 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-Volume 1, pp. 1127–1136. Association for Computational Linguistics (2011)
28. Huang, R., Riloff, E.: Peeling back the layers: Detecting event role fillers in secondary contexts. In: ACL 2011, pp. 1137–1147 (2011)
29. Huff, A.G., Breit, N., Allen, T., Whiting, K., Kiley, C.: Evaluation and verification of the global rapid identification of threats system for infectious diseases in textual data sources. Interdisciplinary perspectives on infectious diseases **2016** (2016)
30. Ide, N., Woolner, D.: Exploiting semantic web technologies for intelligent access to historical documents. In: LREC. Citeseer (2004)

31. Journet, N., Visani, M., Mansencal, B., Van-Cuong, K., Billy, A.: Doccreator: A new software for creating synthetic ground-truthed document images. Journal of imaging **3**(4), 62 (2017)
32. Lai, V., Nguyen, M.V., Kaufman, H., Nguyen, T.H.: Event extraction from historical texts: A new dataset for black rebellions. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 2390–2400. Association for Computational Linguistics, Online (2021). DOI 10.18653/v1/2021.findings-acl.211. URL https://aclanthology.org/2021.findings-acl.211
33. Lejeune, G., Brixtel, R., Doucet, A., Lucas, N.: Multilingual event extraction for epidemic detection. Artificial intelligence in medicine **65** (2015). DOI 10.1016/j.artmed.2015.06.005
34. Lejeune, G., Zhu, L.: A new proposal for evaluating web page cleaning tools. Computacion y Sistemas **22**(4), 1249–1258 (2018)
35. 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 (Volume 1: Long Papers), pp. 73–82. Association for Computational Linguistics, Sofia, Bulgaria (2013). URL https://www.aclweb.org/anthology/P13-1008
36. Linhares Pontes, E., Cabrera-Diego, L.A., Moreno, J.G., Boros, E., Hamdi, A., Sidère, N., Coustaty, M., Doucet, A.: Entity linking for historical documents: Challenges and solutions. In: E. Ishita, N.L.S. Pang, L. Zhou (eds.) Digital Libraries at Times of Massive Societal Transition, pp. 215–231. Springer International Publishing, Cham (2020)
37. Linhares Pontes, E., Hamdi, A., Sidere, N., Doucet, A.: Impact of OCR quality on named entity linking. In: Digital Libraries at the Crossroads of Digital Information for the Future - 21st International Conference on Asia-Pacific Digital Libraries, ICADL 2019, Kuala Lumpur, Malaysia, November 4-7, 2019, Proceedings, pp. 102–115 (2019). DOI 10.1007/978-3-030-34058-2_11. URL https://doi.org/10.1007/978-3-030-34058-2_11
38. 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)
39. 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)
40. Lucas, N.: The enunciative structure of news dispatches, a contrastive rhetorical approach. Language, culture, rhetoric pp. 154–164 (2004)
41. Lucas, N.: Modélisation différentielle du texte, de la linguistique aux algorithmes. Ph.D. thesis, Université de Caen (2009)
42. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: Y. Bengio, Y. LeCun (eds.) 1st International Conference on Learning Representations, ICLR 2013, p. . IEEE, Scottsdale, Arizona, USA (2013)
43. 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, "" (2000)
44. Muller, B., Sagot, B., Seddah, D.: Enhancing bert for lexical normalization. In: Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019), pp. 297–306 (2019)
45. Mutuvi, S., Boros, E., Doucet, A., Lejeune, G., Jatowt, A., Odeo, M.: Multilingual epidemiological text classification: A comparative study. In: COLING, International Conference on Computational Linguistics (2020)
46. Mutuvi, S., Doucet, A., Odeo, M., Jatowt, A.: Evaluating the impact of ocr errors on topic modeling. In: International Conference on Asian Digital Libraries, pp. 3–14. Springer, Berlin, Germany (2018)
47. 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. Association for Computational Linguistics, San Diego, California (2016). DOI 10.18653/v1/N16-1034. URL https://www.aclweb.org/anthology/N16-1034
48. 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 (Volume 2: Short Papers), pp. 365–371. Association for Computational Linguistics, Beijing, China (2015). DOI 10.3115/v1/P15-2060. URL https://www.aclweb.org/anthology/P15-2060
49. Nguyen, T.H., Grishman, R.: Modeling skip-grams for event detection with convolutional neural networks. In: Proceedings of EMNLP (2016)
50. Nguyen, T.T.H., Jatowt, A., Coustaty, M., Nguyen, N.V., Doucet, A.: Deep statistical analysis of ocr errors for effective post-ocr processing. In: Proceedings of the 18th Joint Conference on Digital Libraries, pp. 29–38 (2019)
51. Oberbichler, S., Boroş, 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. Journal of the Association for Information Science and Technology (2021)
52. Pruthi, D., Dhingra, B., Lipton, Z.C.: Combating adversarial misspellings with robust word recognition. In: 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019), pp. 5582–5591. Florence, Italy (2019)
53. Riloff, E.: Automatically generating extraction patterns from untagged text. In: AAAI'96, pp. 1044–1049 (1996)
54. Riloff, E.: An empirical study of automated dictionary construction for information extraction in three domains. Artificial intelligence **85**(1), 101–134 (1996)
55. Rodriquez, K.J., Bryant, M., Blanke, T., Luszczynska, M.: Comparison of named entity recognition tools for raw OCR text. In: J. Jancsary (ed.) 11th Conference on Natural Language Processing, KONVENS 2012, Empirical Methods in Natural Language Processing, September 19-21, 2012, *Scientific series of the ÖGAI*, vol. 5, pp. 410–414. ÖGAI, Wien, Österreich, Vienna, Austria (2012). URL http://www.oegai.at/konvens2012/proceedings/60_rodriquez12w/
56. Rovera, M., Nanni, F., Ponzetto, S.P.: Providing advanced access to historical war memoirs through the identification of events, participants and roles (2019)
57. 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). URL https://aclanthology.org/H05-1088
58. Shaw, R.B.: Events and periods as concepts for organizing historical knowledge. University of California, Berkeley (2010)
59. Smith, R.: An overview of the tesseract ocr engine. In: Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), vol. 2, pp. 629–633. IEEE, IEEE Computer Society, USA (2007)
60. Smith, S.L., Kindermans, P., Le, Q.V.: Don't decay the learning rate, increase the batch size. CoRR **abs/1711.00489**, (2017). URL http://arxiv.org/abs/1711.00489
61. Sprugnoli, R.: Event detection and classification for the digital humanities. Ph.D. thesis, University of Trento (2018)
62. van Strien, D., Beelen, K., Ardanuy, M.C., Hosseini, K., McGillivray, B., Colavizza, G.: Assessing the impact of ocr quality on downstream nlp tasks. ICAART 2020 - Proceedings of the 12th International Conference on Agents and Artificial Intelligence **1**, 484–496 (2020)
63. Sun, L., Hashimoto, K., Yin, W., Asai, A., Li, J., Yu, P., Xiong, C.: Adv-bert: Bert is not robust on misspellings! generating nature adversarial samples on bert. arXiv preprint arXiv:2003.04985 (2020)
64. Ukkonen, E.: Maximal and minimal representations of gapped and non-gapped motifs of a string. Theoretical Computer Science **410**, 4341–4349 (2009). DOI 10.1016/j.tcs.2009.07.015
65. Walker, C., Stephanie, S., Julie, M., Kazuaki, M.: Ace 2005 multilingual training corpus. Technical report, Linguistic Data Consortium (2005)
66. Wang, P., Sun, R., Zhao, H., Yu, K.: A new word language model evaluation metric for character based languages. In: M. Sun, M. Zhang, D. Lin, H. Wang (eds.) Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data, pp. 315–324. Springer Berlin Heidelberg, Berlin, Heidelberg (2013)
67. Yangarber, R., Grishman, R., Tapanainen, P., Huttunen, S.: Automatic acquisition of domain knowledge for information extraction. In: 18th Internation Conference on Computational Linguistics (COLING 2000), pp. 940–946 (2000)