Impact Analysis of Document Digitization on Event Extraction

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

Impact Analysis of Document Digitization on Event Extraction*

Nhu Khoa Nguyen[1][0000−0003−2751−5349], Emanuela Boros[1][0000−0001−6299−9452], Gaël Lejeune[2][0000−0002−4795−2362], and Antoine Doucet[1][0000−0001−6160−3356]

[1] University of La Rochelle, L3i, F-17000, La Rochelle, France
firstname.lastname@univ-lr.fr
https://www.univ-larochelle.fr
[2] Sorbonne University, F-75006 Paris, France
firstname.lastname@sorbonne-universite.fr
https://www.sorbonne-universite.fr/

Abstract. This paper tackles the epidemiological event extraction task applied to digitized documents. Event extraction is an information extraction task that focuses on identifying event mentions from textual data. In the context of event-based health surveillance from digitized documents, several key issues remain challenging in spite of great efforts. First, image documents are indexed through their digitized version and thus, they may contain numerous errors, e.g. misspellings. Second, it is important to address international news, which would imply the inclusion of multilingual data. To clarify these important aspects of how to extract epidemic-related events, it remains necessary to maximize the use of digitized data. In this paper, we investigate the impact of working with digitized multilingual documents with different levels of synthetic noise over the performance of an event extraction system. This type of analysis, to our knowledge, has not been alleviated in previous research.

Keywords: Information Extraction · Event Extraction · Event Detection · Multilingualism.

1 Introduction

The surveillance of epidemic outbreaks has been an ongoing challenge globally and it has been a key component of public health strategy to contain diseases spreading. While digital documents have been the standard format in the modern days, many archives and libraries still keep printed historical documents and records. Historians and geographers have a growing interest in these documents as they still hold many crucial information and events in the past to analyze, noticeably in health and related to epidemics events in an international context.

* This work has been supported by the European Union's Horizon 2020 research and innovation program under grants 770299 (NewsEye) and 825153 (Embeddia).

Event extraction (EE) is an important information extraction (IE) task that focuses on identifying event mentions from text and extracting information relevant to them. Typically, this entails predicting event triggers, the occurrence of events with specific types, and extracting arguments associated with an event. In the context of event-based health surveillance from digitized documents, for extracting relevant events, even though the historical documents are in physical form, few of them have been converted into digital form for further storage as records in a database. However, due to the digitization process, several issues can arise, most commonly in the case when the original document is distorted, whether through deterioration due to aging or was damaged in the storing process, which will affect the converted content. Moreover, errors from the digitization process could also be a factor that causes adulteration of the converted documents e.g. word variations or mispellings.

In this article, we propose to experiment with an approach to event extraction with the ability of handling not only multilingual data, but also large amounts of data without relying on any additional natural language processing (NLP) tools. The architecture is based on the DANIEL system [11] which is a discourse-level approach that exploits the global structure of news. It also tackles the difficulty of language adaptation by its character-based approach that uses positions of substring occurrences in text. We believe that DANIEL is adequate for its ability to handle text in any language and that its algorithm should be robust to noise. We aim at testing the robustness of this model against noise, its ability of treating 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. Furthermore, we create synthetic data starting from the initial dataset in order to study the direct impact of automatic text recognition (ATR) over the performance of both approaches.

The paper is organized as follows: Section 2 briefly overviews the related works on epidemiological event extraction. Section 3 introduces the DANIEL system and its characteristics and in Section 4 the dataset built specifically for the DANIEL system is presented in detail. The Section 5 describes the experiments and an extrinsic evaluation of the results. We conclude and propose possible suggestions for future research in Section 6.

2 Related Work

Specific to epidemiological event extraction, there exist a few of empirical works targeted to extract events related to disease outbreaks. For instance, similar to the chosen system for this paper, DANIEL, there are two other systems, BIOCASTER [2,3] and PULS [5]. These architectures produced adequate results in analyzing disease-related news reports and providing a summary of the epidemics. For example, the BIOCASTER, an ontology-based text mining system, processes and analyzes web texts for the occurrence of disease outbreak in four phases namely, topic classification, named entity recognition (NER), disease/location detection and event extraction.

To our knowledge, there are no works related to the analysis of the impact of documents digitization for event extraction in the epidemiological domain. In return, few studies have been devoted to other information extraction tasks i.e. the extraction of named entities from digitized historical data [1,4]. Dealing with noisy data, several efforts have been devoted to extracting named entities from diverse text types such as outputs of automatic speech recognition (ASR) systems [6,9], informal messages and noisy social network posts [16]. Other researchers [8] quantitatively estimate the impact of digitization quality on the performance of named entity recognition. Other studies focused on named entity linking [15], more specifically on the evaluation of the performance of named entity linking over digitized documents with different levels of digitization quality.

3 Approach

DAnIEL [11] stands for Data Analysis for Information Extraction in any Language. The approach is at document-level, as opposed to the commonly used analysis at sentence-level, by exploiting the global structure of news as defined by the authors of [14]. The entries of the system are news texts, title and body of text, the name of the source when available, and other metadata (e.g date of article). As the name implies, the system has the capability to work in a multilingual setting due to the fact that it is not a word-based algorithm, segmentation in words can be highly language-specific, but rather a character-based one that centers around the repetition and position of character sequences.

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 [7,14], the system is able to detect crucial information in salient zones that are peculiar to this genre of writing: the properties of the journalistic genre and the style universals form the basis of the analysis. This combines with the fact that DAnIEL considers text as sequences of characters, instead of words, the system 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 than often, initial reports where patient zero appears are in the vernacular language. The approach presented in [11] considers the document as the main unit and aims at the language-independent organizational properties that repeat information at explicit locations. According to the author, epidemic news reports, which use the journalistic writing style, have well-defined rules on structure and vocabulary to convey concisely and precisely the message to their targeted audience. As these rules are at a higher level than grammar rules conceptually, they are applied to many languages, thus offer high robustness in a multilingual scenario.

DAnIEL uses a minimal knowledge base for matching between the extracted possible disease names or locations and the knowledge base entries. Its central processing chain includes four phases. In the *Article segmentation* phase, the system first divides the document into salient positions: title, header, body and

footer. In *Pattern extraction*, for detecting events, the system looks for repeated substrings at the salient zones aforementioned. In *Pattern filtering*, the substrings that satisfy this condition will be matched to a list of disease/location names that was constructed by crawling from Wikipedia.

For the string matching between the extracted character sequences and knowledge base entries, the system is parameterized with a ratio. For instance, a small ratio value could offer a perfect recall but with high noise (many irrelevant entries are selected). For a maximum value (1.0), the system will match the exact extracted substrings which could be detrimental to the morphologically rich languages (e.g. Greek, Russian). There are cases where the canonical disease name cannot be found in the text, as in the case of aforementioned languages, but grammatical cases of nouns. For example, in Russian, "Простуда" ("prostuda") means "cold", and since this disease name cannot be found in the text article, we used the instrumental case in Russian that can generally be distinguished by the "-ом" ("-om") suffix for most masculine and neuter nouns, the "-ою/"-ой" ("-oju"/"-oj") suffix for most feminine nouns. A ratio of less than 1.0 will consider the instrumental case for singular "простудой" as a true positive.

Finally, the *Detection of disease − location pairs* (in some cases, the number of victims also) produces the end result with one or more events that are described by pairs of disease-location.

4 Dataset Description

In this section, we present the dataset that was created for the DANIEL system [11]. The corpus is dedicated to multilingual epidemic surveillance and contains articles on different press threads in the field of *health* (Google News) that focused on epidemic events from different collected documents in different languages, with events simply defined as disease-location-number of victims triplets. The corpus was built specifically for this system [11,12], containing articles from six different languages: English, French, Greek, Russian, Chinese, and Polish. It contains articles on different press threads in the field of *health* (Google News) focused on epidemic events and it was annotated by native speakers.

A DANIEL event is defined at document-level, meaning that an article is considered as relevant if it is annotated with a disease − location pairs (and rarely, the number of victims). An example is presented in Figure 1, where the event is a *listeria* outbreak in *USA* and the number of victims is unknown.

Thus, in this dataset the event extraction task is defined as identifying articles that contain an event and the extraction of the disease name and location, i.e. the words or compound words that evoke the event. Since the events are epidemic outbreaks, there is no pre-set list of types and subtypes of events, and thus the task of event extraction is simplified to detecting whether an article contains an epidemiological event or not.

Common to event extraction, the dataset is characterized by imbalance. In this case, only around 10% of these documents are relevant to epidemic events, which is very sparse. The number of documents in each language is rather bal-

Fig. 1. Example of an event annotated in DANIEL dataset.


```
"15960": {
    "annotations": [
        [    "listeria",
             "USA",
             "unknown"]
    ],
    "comment": "",
    "date": "2012-01-12",
    "language": "en",
    "document_path": "doc_en/20120112_www.cnn.
        com_48eddc7c17447b70075c26a1a3b168243edcbfb28f0185",
    "url": "http://www.cnn.com/2012/01/11/health/listeria-
        outbreak/index.html"
}
```


anced, except for French, having about five times more documents compared to the rest of the languages. More statistics on the corpus can be found in Table 1.

The DANIEL dataset is annotated at document-level, which differentiates itself from other datasets used in research for the event extraction task. A document is either reporting an event (disease-place pair, and sometimes the number of victims) or not. We will elaborate the evaluation framework in Section 5.

Table 1. Summary of the DANIEL dataset. The relevant documents are documents annotated with an event.

Language	**# Documents**	**# Relevant**	**# Sentences**	**# Tokens**
French (fr)	2,733	340 (12.44%)	75,461	2,082,827
English (en)	475	31 (6.53%)	4,153	262,959
Chinese (zh)	446	16 (3.59%)	4,555	236,707
Russian (ru)	426	41 (9.62%)	6,865	133,905
Greek (el)	390	26 (6.67%)	3,743	198,077
Polish (pl)	352	30 (8.52%)	5,847	165,682
Total	4,822	489 (10.14%)	140,624	3,080,157

5 Experiments

The focal point of this set of experiments is to observe how the level of noise stemming from the digitization process impacts the performance of the models. However, there is no adequate historical document dataset provided with manually curated event annotation that could directly be used to measure the per-

formance of the models over deteriorated historical documents. Thus, the noise and degradation levels have to be artificially generated into clean documents, so as to measure the impact of ATR over event detection using DAnIEL. We shall thus use readily available data sets over contemporary and digitally-born datasets, which are free of any ATR-induced noise.

In order to create such an appropriate dataset, the raw text from the DANIEL dataset was extracted and converted into clean images[3]. 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. We used four types of noise: *Character Degradation* adds small ink dots on characters to emulate the age effect on articles, *Phantom Character* appears when characters erode due to excessive use of documents, *Bleed Through* appears in double-paged document image scans where the content of the back side appears in the front side as interference, and *Blur* is a common degradation effect encountered during a typical digitization process. After contaminating the corpus, all the text was extracted from noisy images[4], for initial clean images (without any adulteration) and the noisy synthetic ones. An example with the degradation levels is illustrated in Figure 2. The noise levels were empirically chosen with a considerable level of difficulty[5].

The experiments were conducted in the following manner: for each noise type, the different intensity is generated to see its relation to the performance of the model. Character error rate (CER) and word error rate (WER) were calculated for each noise level, that can align long noisy text even with additional or missing text with the ground truth, thus enables it to calculate the error rate of OCR process. The experiments are performed under conditions of varying word error rate (WER) and character error rate (CER): original text, OCR from high-quality text images, and OCR on synthetically degraded text images.

5.1 Evaluation Framework

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 (CER) and word error rate (WER). We perform two types of evaluations, both at the document level (included in the DANIEL system):

- Event identification: a document represents an event if both triggers were found, regardless of their types;
- Event classification: a document represents an event if the triggers are correctly found and match exactly with the groundtruth ones.

[3] For simulating different levels of degradation, we used DocCreator [10].

[4] The Tesseract optical character recognition (OCR) Engine v4.0 `https://github.com/tesseract-ocr/tesseract` [17] was used to produce the digitised documents.

[5] The following values of DocCreator are: *Character Degradation* (2-6), *Phantom Character* (Very Frequent), *Blur* (1-3), *Bleed Through* (80-80).

Fig. 2. Example of types of noise applied on a dataset: (i) clean image, (ii) *Phantom Character*, (iii) *Character Degradation*, (iv)*Bleed Through*, (v) *Blur*, and (vi) all mixed together.

5.2 Experiments with Clean Data

Hereafter, we present the experiments performed with the clean data. Considering that the DANIEL system has a ratio parameter for matching the extracted triggers, we test two values for it. For the first experiments, we use a ratio value of 0.8 (the default value of the system) that was empirically chosen in [11] for the best trade-off between recall and precision. Second, we test the maximum ratio value of 1.0 in order to analyze the system's performance when the extracted disease names and locations exactly match with the knowledge base.

For event identification on clean textual data, one can notice from the Table 2, that usually DANIEL favors 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. It is aso not surprising that the DANIEL system the highest performance values for event identification for Chinese and Greek, since for Chinese, there are few relevant documents comparing with the other languages (16 documents that report an event), and for Greek, there are 26 of them.

Table 2. Evaluation of DANIEL on the initial dataset for event identification (regardless of the types of the triggers).

		Polish	Chinese	Russian	Greek	French	English	All languages
ratio=0.8	P	0.6842	0.8	0.7115	0.641	0.592	0.4918	0.6052
	R	0.8667	1.0	0.9024	0.9259	0.9088	0.8571	0.9059
	F1	0.7647	0.8889	0.7957	0.7576	0.7169	0.625	0.7256
ratio=1.0	P	0.0	0.0	0.0	0.0	0.9155	0.0	0.9155
	R	0.0	0.0	0.0	0.0	0.5735	0.0	0.3988
	F1	0.0	0.0	0.0	0.0	0.7052	0.0	0.5556

We also can note the large difference between the two chosen ratios. More exactly, an increase in this value comes in the detriment of the languages that are not only morphologically rich, but also in the case where the exact name of the disease is not located in the text.

Table 3. Evaluation of DANIEL for event classification (triggers are correctly found and match with the groundtruth ones).

		Polish	Chinese	Russian	Greek	French	English	All languages
ratio=0.8	P	0.3421	0.35	0.2692	0.4103	0.5211	0.2951	0.4645
	R	0.4	0.4118	0.3146	0.5079	0.5781	0.4737	0.5363
	F1	0.3688	0.3784	0.2902	0.4539	0.5481	0.3636	0.4978
ratio=1.0	P	0.0	0.0	0.0	0.0	0.7934	0.0	0.7934
	R	0.0	0.0	0.0	0.0	0.3592	0.0	0.2666
	F1	0.0	0.0	0.0	0.0	0.4945	0.0	0.3991

In the case of event classification, we observe from Table 3, that DANIEL is balanced regarding the precision and recall metrics, being able to have higher F1 on the under-represented languages (Chinese, Russian, and Greek). We also notice that, in all the cases, DANIEL 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, e.g. 10000 cannot be detected since the original text has the $10,000$ form, or it is spelled *ten thousands*. Generally, for the detection of locations, we recall that DAnIEL is capable to detect locations due to the usage of external resources and article metadata.

For the experiments on noisy data, we will use a ratio value of 0.8, since the maximum value for the ratio creates results prone to suffer from word variations or misspelings of words (which is a direct consequence of the digitization process).

5.3 Experiments with Noisy Data

The results in Table 4 clearly state that *Character Degradation* is the effect that affects the most the transcription of the documents. However, for character-

Table 4. Document degradation OCR evaluation on the DANIEL dataset.

		Clean	CharDeg	Bleed	Blur	Phantom	All
All	CER	2.61	9.55	2.83	8.76	2.65	11.07
	WER	4.23	26.23	5.93	19.05	4.71	27.36
Polish	CER	0.15	5.86	0.19	7.57	0.19	5.51
	WER	0.74	20.66	1.17	13.23	1.17	20.70
Chinese	CER	36.89	41.01	38.24	43.97	36.91	46.97
	WER	–	–	–	–	–	–
Russian	CER	0.93	16.20	1.45	8.13	1.03	10.91
	WER	1.63	28.46	6.61	14.94	2.73	29.72
Greek	CER	3.52	9.04	3.76	13.79	3.54	16.28
	WER	15.86	41.36	17.39	54.02	15.93	54.76
French	CER	1.96	8.37	2.13	7.43	2.0	10.90
	WER	3.33	23.56	4.89	16.31	3.76	26.07
English	CER	0.35	5.75	0.52	4.74	0.44	7.43
	WER	0.66	24.78	2.14	14.72	1.66	20.99

based languages (e.g. Chinese), CER is commonly used instead of WER as the measure for OCR, and, thus, we report only the CER [18].

We note also that, regarding the Chinese documents, the high values for CER, for every type of noise, might be caused by the existence of the enormous number of characters in the alphabet that, by adding such an effect as *Character Degradation* can change drastically the recognition of a character (and in Chinese, one single character can often be a word). Otherwise, while *Character Degradation* noise and *Blur* effect have more impact on the performance of DANIEL than *Phantom Character* type since it did not generate enough distortion to the images. A similar case applies for the *Bleed Through* noise.

Regarding the experiments presented in Tables 5 and 6, we notice, first of all, that the *Character Degradation* effect, *Blur*, and most of all, all the effects mixed together, have indeed an impact or effect over the performance of DAnIEL, but with little variability. Meanwhile, *Phantom Degradation* and *Bleed through* had very little to no impact on the quality of detection with DAnIEL.

The cause of the decrease in performance of DANIEL 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 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.

Regarding all the aforementioned results for the DANIEL system, computing the number of affected event words (disease, location, number of cases), 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, close to the 1.63% that were affected by the OCR on clean data. This is due to the imbalance in the DANIEL dataset: only 10.14% of a total of 4,822 documents

Table 5. Evaluation of DANIEL results on the noisy data for event identification (regardless of the types of the triggers). Orig=Original, PL=Polish, ZH=Chinese, RU=Russian, EL=Greek, FR=French, EN=English.

		Orig	Clean	CharDeg	Bleed	Blur	Phantom	All
All	P	0.61	0.735 (+0.12)	0.755 (+0.14)	0.735 (+0.12)	0.74 (+0.13)	0.731 (+0.12)	0.758 (+0.14)
	R	0.91	0.859 (-0.05)	0.674 (-0.23)	0.862 (-0.04)	0.857 (-0.05)	0.862 (-0.04)	0.718 (-0.19)
	F1	0.73	0.792 (+0.06)	0.712 (-0.01)	0.793 (+0.06)	0.794 (+0.06)	0.791 (+0.06)	0.737 (+0.00)
PL	P	0.68	0.643 (-0.03)	0.656 (-0.02)	0.658 (-0.02)	0.692 (+0.01)	0.643 (-0.03)	0.645 (-0.03)
	R	0.87	0.9 (+0.03)	0.7 (-0.17)	0.9 (+0.03)	0.9 (+0.03)	0.9 (+0.03)	0.667 (-0.20)
	F1	0.76	0.75 (-0.01)	0.677 (-0.08)	0.761 (+0.00)	0.783 (+0.02)	0.75 (-0.01)	0.656 (-0.10)
ZH	P	0.8	0.882 (+0.08)	0.882 (+0.08)	0.789 (-0.01)	0.733 (-0.06)	0.789 (-0.01)	0.857 (+0.05)
	R	1.0	0.938 (-0.06)	0.938 (-0.06)	0.938 (-0.06)	0.917 (-0.08)	0.938 (-0.06)	0.75 (-0.25)
	F1	0.89	0.909 (+0.01)	0.909 (+0.01)	0.857 (-0.03)	0.815 (-0.07)	0.857 (-0.03)	0.8 (-0.09)
RU	P	0.71	0.688 (-0.02)	0.691 (-0.01)	0.688 (-0.02)	0.705 (-0.00)	0.688 (-0.02)	0.727 (+0.01)
	R	0.9	0.805 (-0.09)	0.744 (-0.15)	0.846 (-0.05)	0.795 (-0.10)	0.846 (-0.05)	0.821 (-0.08)
	F1	0.8	0.742 (-0.05)	0.716 (-0.08)	0.759 (-0.04)	0.747 (-0.05)	0.759 (-0.04)	0.771 (-0.02)
EL	P	0.64	0.59 (-0.05)	0.682 (+0.04)	0.59 (-0.05)	0.639 (-0.00)	0.59 (-0.05)	0.667 (+0.02)
	R	0.93	0.852 (-0.07)	0.556 (-0.37)	0.852 (-0.07)	0.852 (-0.07)	0.852 (-0.07)	0.518 (-0.41)
	F1	0.76	0.697 (-0.06)	0.612 (-0.14)	0.697 (-0.06)	0.73 (-0.03)	0.697 (-0.06)	0.583 (-0.17)
FR	P	0.59	0.803 (+0.21)	0.828 (+0.23)	0.806 (+0.21)	0.801 (+0.21)	0.801 (+0.21)	0.816 (+0.22)
	R	0.91	0.849 (-0.06)	0.666 (-0.24)	0.849 (-0.06)	0.849 (-0.06)	0.849 (-0.06)	0.723 (-0.18)
	F1	0.72	0.826 (+0.10)	0.738 (+0.01)	0.827 (+0.10)	0.825 (+0.10)	0.825 (+0.10)	0.767 (+0.04)
EN	P	0.49	0.508 (+0.01)	0.458 (-0.03)	0.508 (+0.01)	0.516 (+0.02)	0.508 (+0.01)	0.52 (+0.03)
	R	0.86	0.943 (+0.08)	0.629 (-0.23)	0.943 (+0.08)	0.943 (+0.08)	0.943 (+0.08)	0.743 (-0.11)
	F1	0.62	0.66 (+0.04)	0.53 (-0.09)	0.66 (+0.04)	0.667 (+0.04)	0.66 (+0.04)	0.612 (-0.00)

contain events. It brings us to the conclusion that the event extraction task is not considerably impacted by the degradation of the image documents.

One interesting observation is that the precision or the recall can increase, resulting in a higher F1, despite the higher noise effect applied. One possible explanation for this phenomenon is that with a greater level of noise, some false positives disappear. Documents, which were previously classified wrongly due to being too ambiguous to the system (for instance documents relating vaccination campaigns are usually tagged as non-relevant in the ground truth dataset), were given much more distinction thanks to the noise, thus making them look less like relevant samples to the system. More formally: let document X be a false positive in its raw format (X_{raw}). Let X_{Noisy} be its noisy version. If the paragraph that triggered both system's misclassifications disappeared in X_{noisy}, there are good chances that it will be classified as non-relevant. In that case, X_{raw} is a false positive but X_{noisy} is a true negative. That may seem counter-intuitive but noise can improve classification results, see for instance [13] for a study on the same dataset of the influence of boilerplate removal on results.

Table 6. Evaluation of DANIEL results on the noisy data for event classification (triggers are correctly found and match with the groundtruth ones). Orig=Original, PL=Polish, ZH=Chinese, RU=Russian, EL=Greek, FR=French, EN=English.

		Orig	Clean	CharDeg	Bleed	Blur	Phantom	All
All	P	0.46	0.552 (+0.09)	0.548 (+0.08)	0.549 (+0.08)	0.558 (+0.09)	0.548 (+0.08)	0.547 (+0.08)
	R	0.54	0.497 (-0.04)	0.377 (-0.16)	0.496 (-0.04)	0.497 (-0.04)	0.498 (-0.04)	0.4 (-0.14)
	F1	0.5	0.523 (+0.02)	0.447 (-0.05)	0.521 (+0.02)	0.526 (+0.02)	0.521 (+0.02)	0.462 (-0.03)
PL	P	0.34	0.333 (-0.00)	0.328 (-0.01)	0.342 (+0.00)	0.359 (+0.01)	0.333 (-0.00)	0.274 (-0.06)
	R	0.4	0.431 (+0.03)	0.323 (-0.07)	0.431 (+0.03)	0.431 (+0.03)	0.431 (+0.03)	0.262 (-0.13)
	F1	0.37	0.376 (+0.00)	0.326 (-0.04)	0.381 (+0.01)	0.392 (+0.02)	0.376 (+0.00)	0.268 (-0.10)
ZH	P	0.35	0.412 (+0.06)	0.353 (+0.00)	0.342 (-0.00)	0.367 (+0.01)	0.342 (-0.00)	0.464 (+0.11)
	R	0.41	0.412 (+0.00)	0.353 (-0.05)	0.382 (-0.02)	0.423 (+0.01)	0.382 (-0.02)	0.382 (-0.02)
	F1	0.38	0.412 (+0.03)	0.353 (-0.02)	0.361 (-0.01)	0.393 (+0.01)	0.361 (-0.01)	0.419 (+0.03)
RU	P	0.27	0.302 (+0.03)	0.312 (+0.04)	0.302 (+0.03)	0.295 (+0.02)	0.302 (+0.03)	0.273 (+0.00)
	R	0.31	0.326 (+0.01)	0.357 (+0.04)	0.341 (+0.03)	0.306 (-0.00)	0.341 (+0.03)	0.282 (-0.02)
	F1	0.31	0.314 (+0.00)	0.333 (+0.02)	0.32 (+0.01)	0.301 (-0.00)	0.32 (+0.01)	0.278 (-0.03)
EL	P	0.41	0.333 (-0.07)	0.341 (-0.06)	0.333 (-0.07)	0.361 (-0.04)	0.333 (-0.07)	0.357 (-0.05)
	R	0.51	0.413 (-0.09)	0.238 (-0.27)	0.413 (-0.09)	0.413 (-0.09)	0.413 (-0.09)	0.238 (-0.27)
	F1	0.45	0.369 (-0.08)	0.28 (-0.17)	0.369 (-0.08)	0.385 (-0.06)	0.369 (-0.08)	0.286 (-0.16)
FR	P	0.47	0.691 (+0.22)	0.693 (+0.22)	0.69 (+0.22)	0.689 (+0.21)	0.689 (+0.21)	0.675 (+0.20)
	R	0.51	0.527 (+0.01)	0.402 (-0.10)	0.524 (+0.01)	0.527 (+0.01)	0.527 (+0.01)	0.431 (-0.07)
	F1	0.49	0.598 (+0.10)	0.509 (+0.01)	0.596 (+0.10)	0.597 (+0.10)	0.597 (+0.10)	0.526 (+0.03)
EN	P	0.47	0.292 (-0.17)	0.26 (-0.21)	0.292 (-0.17)	0.297 (-0.17)	0.292 (-0.17)	0.31 (-0.16)
	R	0.51	0.5 (-0.01)	0.329 (-0.18)	0.5 (-0.01)	0.5 (-0.01)	0.5 (-0.01)	0.408 (-0.10)
	F1	0.49	0.369 (-0.12)	0.291 (-0.19)	0.369 (-0.12)	0.372 (-0.11)	0.369 (-0.12)	0.352 (-0.13)

6 Conclusions and Perspectives

We conclude that, in our experimental setting, the epidemical event extraction is prone to digitization errors, but, at the same time, the impact on the DANIEL system is not considerable, which makes it a robust solution for health surveillance applications. Nevertheless, while these experiments were performed in an artificial setting with synthetically produced noise effects, the challenges that exist in a more realistic reasonable scenario could generate other tremendous issues due to the digitization process. As a perspective, we consider the annotation of a digitized dataset in order to asses our assumptions.

References

1. Byrne, K.: Nested named entity recognition in historical archive text. In: International Conference on Semantic Computing (ICSC 2007). pp. 589–596. IEEE (2007)
2. Collier, N.: Towards cross-lingual alerting for bursty epidemic events. Journal of Biomedical Semantics **2**(5), S10 (2011)
3. 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)

4. Crane, G., Jones, A.: The challenge of virginia banks: an evaluation of named entity analysis in a 19th-century newspaper collection. In: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries. pp. 31–40 (2006)
5. Du, M., Von Etter, P., Kopotev, M., Novikov, M., Tarbeeva, N., Yangarber, R.: Building support tools for russian-language information extraction. In: International Conference on Text, Speech and Dialogue. pp. 380–387. Springer (2011)
6. Favre, B., Béchet, F., Nocéra, P.: Robust named entity extraction from large spoken archives. In: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. pp. 491–498. Association for Computational Linguistics (2005)
7. 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 (03 2018). https://doi.org/10.1007/978-3-319-78105-1_39
8. Hamdi, A., Jean-Caurant, A., Sidère, N., Coustaty, M., Doucet, A.: Assessing and minimizing the impact of ocr quality on named entity recognition. In: International Conference on Theory and Practice of Digital Libraries. pp. 87–101. Springer (2020)
9. Hatmi, M.: Reconnaissance des entités nommées dans des documents multimodaux. Ph.D. thesis, Université de Nantes (2014)
10. 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)
11. Lejeune, G., Brixtel, R., Doucet, A., Lucas, N.: Multilingual event extraction for epidemic detection. Artificial intelligence in medicine **65** (07 2015). https://doi.org/10.1016/j.artmed.2015.06.005
12. Lejeune, G., Doucet, A., Yangarber, R., Lucas, N.: Filtering news for epidemic surveillance: towards processing more languages with fewer resources. In: Proceedings of the 4th Workshop on Cross Lingual Information Access. pp. 3–10 (2010)
13. Lejeune, G., Zhu, L.: A new proposal for evaluating web page cleaning tools. Computacion y Sistemas **22**(4), 1249–1258 (2018)
14. Lucas, N.: Modélisation différentielle du texte, de la linguistique aux algorithmes. Ph.D. thesis, Université de Caen (2009)
15. Pontes, E.L., Hamdi, A., Sidere, N., Doucet, A.: Impact of ocr quality on named entity linking. In: International Conference on Asian Digital Libraries. pp. 102–115. Springer (2019)
16. Ritter, A., Clark, S., Etzioni, O., et al.: Named entity recognition in tweets: an experimental study. In: Proceedings of the conference on empirical methods in natural language processing. pp. 1524–1534. Association for Computational Linguistics (2011)
17. 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 (2007)
18. Wang, P., Sun, R., Zhao, H., Yu, K.: A new word language model evaluation metric for character based languages. In: Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data, pp. 315–324. Springer (2013)