Novel Event Detection and Classification for Historical Texts

Rachele Sprugnoli
Università Cattolica del Sacro Cuore
Linguistic Sciences and
Foreign Literature Department
rachele.sprugnoli@unicatt.it

Sara Tonelli
Fondazione Bruno Kessler
Digital Humanities Research Group
satonelli@fbk.eu

Event processing is an active area of research in the Natural Language Processing community, but resources and automatic systems developed so far have mainly addressed contemporary texts. However, the recognition and elaboration of events is a crucial step when dealing with historical texts Particularly in the current era of massive digitization of historical sources: Research in this domain can lead to the development of methodologies and tools that can assist historians in enhancing their work, while having an impact also on the field of Natural Language Processing. Our work aims at shedding light on the complex concept of events when dealing with historical texts. More specifically, we introduce new annotation guidelines for event mentions and types, categorized into 22 classes. Then, we annotate a historical corpus accordingly, and compare two approaches for automatic event detection and classification following this novel scheme. We believe that this work can foster research in a field of inquiry as yet underestimated in the area of Temporal Information Processing. To this end, we release new annotation guidelines, a corpus, and new models for automatic annotation.

1. Introduction

In the last decade, the process of searching, understanding, organizing, and synthesizing the content of historical documents has been influenced by the ever-increasing proliferation of digitized sources over the Web (Bingham 2010). While in the 1990s major efforts were devoted to the production of digital pictures by scanning manuscripts and books, the availability of searchable texts in online databases, digital archives, and digital libraries exponentially grew over the 2000s thanks to the development of computer techniques such as Optical Character Recognition. Projects like the Perseus Digital Library[1] are facilitating the access to documents, but, on the other hand, in such a vast quantity of online material that even expert users risk getting lost.

1 http://www.perseus.tufts.edu/.

Submission received: 8 June 2018; revised version received: 29 December 2018; accepted for publication: 10 February 2019.

doi:10.1162/COLI_a_00347

Because events are commonly considered as the building blocks of historical knowledge with which historians construct their system of ideas about the past (Oakeshott 2015; Shaw 2010), a systematic and consistent analysis of events mentioned in historical texts would greatly contribute to a better understanding of large archives in this domain. Furthermore, the extensive work done in the Natural Language Processing (NLP) community to manually annotate and automatically detect event mentions could represent a valuable starting point for this kind of investigation, taking advantage of existing systems for event detection (Llorens, Saquete, and Navarro 2010; Atefeh and Khreich 2015; Reimers and Gurevych 2015; Derczynski et al. 2016) and of the results of past evaluation campaigns (Verhagen et al. 2007; UzZaman et al. 2013; Minard et al. 2015). This cross-fertilization has already been applied to the clinical, chemical, and biomedical domains also for tasks other than event processing (Nédellec et al. 2006; Teufel, Siddharthan, and Batchelor 2009; Sohn et al. 2010), but only limited efforts have been devoted to the historical domain (Cybulska and Vossen 2010; Fokkens et al. 2014).

In this work, we present a novel contribution to analyze and automatically classify event mentions in historical documents. In particular, we develop annotation guidelines and an annotated corpus for the domain, and then we present some experiments for the automatic detection and classification of event mentions in historical texts. This study is built upon the findings in Sprugnoli and Tonelli (2017), where we performed an analysis of historians' requirements regarding linguistic annotation of events in text. Based on the outcome of that previous survey, we present in this work the following contributions:

- new annotation guidelines for the detection and classification of event mentions specifically designed for historical texts;
- a novel corpus of historical texts annotated with events made available to the research community;
- release of word embeddings pre-trained on a corpus of historical texts and of models for the automatic annotation of events developed using two different approaches: traditional linear-chain Conditional Random Fields and a neural architecture.

Detailed guidelines, annotated corpus, pre-trained historical embeddings, and best models are available on GitHub: `https://github.com/dhfbk/Histo`. Further information on the background work that led us to the development of these resources can be found in Sprugnoli (2018).

This article is structured as follows: In Section 2 we provide an overview of how events have been defined in NLP with a focus on the efforts, undertaken in the area of Information Extraction in the last decades, that were directed toward both their manual annotation and their automatic detection and processing. Section 3 presents our guidelines for event mention annotation in historical texts, based on the outcome of the investigation presented in Sprugnoli and Tonelli (2017). In Section 4 we provide details on a newly created corpus annotated following our guidelines: Information on inter-annotator agreement is provided together with an analysis of the final annotated data set. Section 5 describes our experiments aimed at the development of an automatic system for event detection and classification: both a conditional random fields classifier and a neural architecture are tested. Results are compared and discussed in Section 6. Finally, Section 7 provides a summary of the paper and discusses the lessons learned from this research work.

2. Related Work

Most existing works on event extraction have focused on contemporary texts in domains such as news, biomedicine, and medicine (Filatova and Hovy 2001; Katz and Arosio 2001; Schilder and Habel 2003; Björne and Salakoski 2011; Bethard et al. 2015).

Research in the aforementioned domains has been fostered by the organization of many evaluation campaigns and shared tasks, which revised the notion of event to tailor it to the domain of interest. For example, news are the focus of the "Scenario Template" task within the Message Understanding Conference (MUC) (Sundheim 1991) and of the Automatic Content Extraction (ACE), TempEval and the Text Analysis Conference (TAC) Event Nugget series (Doddington et al. 2004; Verhagen et al. 2007, 2010; UzZaman et al. 2013; Mitamura, Liu, and Hovy 2016, 2017). The i2b2 challenge and Clinical-TempEval are instead about the processing of clinical documents (Sun, Rumshisky, and Uzuner 2013; Bethard et al. 2015, 2016, 2017), whereas the Bio-NLP campaign has tasks on event processing in the domain of biomedicine (Kim et al. 2009, 2011; Nédellec et al. 2013; Sun, Rumshisky, and Uzuner 2013; Kim et al. 2016). Within these initiatives, several corpora have been annotated following annotation guidelines subsuming different event definitions and, as a consequence, proposing different mark-up rules in terms of event taggability, linguistic realization, extent and classification.[2]

The most used scheme for the annotation of temporal information is TimeML (Pustejovsky et al. 2003), which is at the basis of the TempEval evaluation campaigns and was also consolidated as an ISO-standard (Iso, SemAf/Time Working Group 2008). For TimeML, events can be linguistically realized by many expressions corresponding to different parts of speech (PoS) such as verbs, nouns, and adjectives in predicate positions. The extension of events must be as short as possible and their classification is based on both their aspectual type (*Aktionsart*) (Vendler 1957) and the syntactic structure they appear in. For example, the `I_STATE` class includes events that describe stative situations and that introduce another event as their argument, like the event *afraid* in *They were afraid to stay*. A minimal extension rule is followed also by the THYME guidelines, for which only the syntactic heads should be annotated for all those events that are "relevant to the patient's clinical timeline" (Styler et al. 2014).

Also, ACE annotates only trigger words encoding an event, whose full extent would be the entire sentence in which it occurs. Trigger words are single tokens, with the only exception of verb+particle continuous constructions (e.g., *laid off*), and discontinuous extensions are not allowed. In ACE not all events are annotated but only those belonging to a predefined set of eight semantic classes, each divided in sub-classes. The same event classification is adopted, with few modifications, also by the ERE (Entities, Relations, Events) scheme and in the Event Nugget annotation (Aguilar et al. 2014; Mitamura et al. 2015). In the latter, the rule is to annotate the maximum extent of text encoding an event, allowing the annotation of continuous and discontinuous multi-token expressions.

As for event annotation in biomedical texts, a domain-specific classification has been proposed in which events involving one or more biological entities are classified following a hierarchical ontology and annotated by experts in the GENIA corpus (Kim et al. 2006). In the current work, our goal is to take advantage of existing practices for event annotation in specialized domains, modifying annotation guidelines when needed to fit the requirements of the history domain. For example, while event extent and realization

2 A list of corpora annotated with events in different languages and related to different domains is available in Sprugnoli and Tonelli (2017).

have been selected with the help of domain experts after looking at existing guidelines, event types were manually adapted from a historical thesaurus, since the available annotation schemes were not able to capture the semantic categories relevant to the domain.

Regarding the development of automatic annotation systems dealing with the schemes previously described, machine learning algorithms and rule-based approaches have been mainly applied in the past, whereas deep learning architectures are more recent. For example, in TempEval 2013 one participanting system adopted a rule-based approach in the event extraction and classification subtask (Zavarella and Tanev 2013), whereas the others were based on Conditional Random Fields (CRFs), MaxEnt classification, and Logistic Regression, taking advantage of morphosyntactic and semantic features (Llorens, Saquete, and Navarro 2010; Jung and Stent 2013; Kolomiyets and Moens 2013). Best results in event extraction are around 0.80 F1-score: results drop significantly in class assignment, with the best score below 0.72. Support Vector Machine (SVM) is the machine learning approach used by the best systems in the Bio-NLP 2016 biomedical event extraction task (Li, Rao, and Zhang 2016; Lever and Jones 2016). A deep learning approach has been proposed too, being ranked second in the identification of localization events of bacteria (Mehryary et al. 2016).

In the clinical domain, the BluLab was the best performing system in the event expression task of Clinical TempEval 2015: It uses SVM algorithms and linguistic features obtaining an F-score of 0.87 on span identification and of 0.82 on class assignment (Velupillai et al. 2015). In 2016, the UTHealth system adopted SVM but added more features, embeddings, and information from domain-specific dictionaries, achieving an F-score of 0.93 on span identification and 0.88 on class assignment (Lee et al. 2016). In 2017, the focus of Clinical TempEval has shifted toward domain adaptation: Systems were trained on a clinical condition (colon cancer data) and tested on another clinical condition (brain cancer data) using an unsupervised approach and also a supervised one but with a limited quantity of training in-domain data. The best system, LIMSI-COT, proposed a deep learning approach for event detection using long short-term memory networks (LSTMs) and a linear SVM for classification. The system obtained an F-score of around 0.70 in the unsupervised setting for both span identification and class assignment and around 0.75 in the supervised one, showing a consistent drop in performance with respect to the previous year (Tourille et al. 2017).

The shift toward deep learning approaches is particularly evident in the 2017 TAC KBP Event Nugget track: 8 out of 10 participating systems used neural network models, both sequential and convolutional. The best performing system is made of an ensemble model with a bidirectional long short-term memory network (BiLSTM) and a CRF. Results range between 0.44 and 0.67 in terms of F1 for event identification and between 0.33 and 0.56 for event classification. In the current work, we choose to compare two approaches, i.e., CRF and the more recent BiLSTM, given its promising performance in the above-mentioned evaluations.

As for the application of NLP to the history domain, only few works were carried out by the community working on Temporal Information Processing. Studies in this field have focused more on the modeling of historical events through the development and use of ontologies (Raimond and Abdallah 2007; Shaw, Troncy, and Hardman 2009; Van Hage et al. 2011; The European Union 2012; Le Boeuf et al. 2017) without fully exploiting NLP methods. Past approaches to event extraction from historical texts have dealt only with verbal events or named events (Ide and Woolner 2007; Segers et al. 2011). Other efforts have been directed to the identification of specific types of events, such as conflict or communication events (Ide and Woolner 2004; Cybulska and Vossen 2011). Finally, more recent works have adopted crowdsourcing techniques or a

frame-based approach (De Boer et al. 2015; Fokkens et al. 2018). Additionally, data annotated with events in this domain and publicly released are scarce: Exceptions are the ModeS TimeBank (Guerrero Nieto, Saurì, and Bernabé Poveda 2011) with Spanish texts from the eighteenth century and the De Gasperi corpus (Speranza and Sprugnoli 2018), a collection of documents written by the Italian statesman Alcide de Gasperi at the beginning of the twentieth century. Both these corpora were annotated following the TimeML guidelines without any adaptation to the history domain. In other words, no attempt was made to find a domain-specific definition of event, unlike what happened in other domains, such as the clinical one. This is the main motivation leading to the current investigation.

3. Events in Historical Texts: Annotation Guidelines

In Sprugnoli and Tonelli (2017) we presented an effort to gather from domain experts requirements about the linguistic annotation of events in the historical domain. This previous work suggested that the development of annotation guidelines for the analysis of texts in a specific domain must be carried out jointly with experts of such a domain. In the case of history scholars, the main findings were that i) the semantic type of an event is the most relevant information to be annotated in historical documents, ii) multi-token annotation of event mentions should be admitted, and iii) events can have different syntactic realizations and grammatical classes.

Taking into account this previous study, we have developed novel annotation guidelines focused on the identification and classification of event mentions. We have adopted a wide definition of event by referring to the notion of eventuality,[3] introduced by Bach (Bach 2008) and re-elaborated by Dölling (Dölling 2014). For this reason, we take into consideration event mentions denoting all types of (punctual or durative) actions, processes, and states. Furthermore, we assume that events can be realized with different parts of speech and syntactic constructions.

Because historical texts are a rather general category spanning diverse topics and genres, we put particular effort into developing a set of semantic classes that offer an exhaustive categorization of events, avoiding too much granularity for annotation purposes but also ensuring informativeness. This led to the definition of 22 semantic classes, thus overcoming the limited classifications proposed by other initiatives such as ACE (Linguistic Data Consortium 2005) and Rich ERE (Song et al. 2015).[4] We summarize below the guidelines developed for the annotation, including information on event extent, linguistic realization, and types. The complete version of the guidelines is available at this link: `https://github.com/dhfbk/Histo/blob/master/Guidelines.pdf`.

3.1 Event Linguistic Realization

In our annotation scheme for historical texts, the linguistic elements that may realize an event are the following:

- verbs in both finite and non-finite form;

  (1) *she **expected** to be **attacked***

3 Hereafter we will use the terms "event" and "eventuality" interchangeably.

4 For a detailed comparison of event annotation in ACE, ERE, and other analogous initiatives, please see Aguilar et al. (2014).

- past participles in the nominal pre-modifier position that represent resultative events. Interpreted as a state, the following example can be paraphrased as "the state of having been imprisoned";

  (2) *an* ***imprisoned*** *criminal*

- present participles in the nominal pre-modifier position that represent in-progress events. In the following example, the modifier describes an event in progress so that it can be paraphrased as "the audience that is smiling and applauding";

  (3) *a* ***smiling*** *and* ***applauding*** *audience*

- adjectives in predicative position;

  (4) *the museum itself* ***was damp***

- nouns that can realize eventualities in different ways:

  – deverbal nouns denoting an activity or an action;

    (5) *the* ***running*** *of these ferries*

  – nouns that have an eventive meaning in their lexical properties even if they do not derive from verbs;

    (6) *delegates of Russia against the* ***war***

  – post-copular nouns;

    (7) *it* ***was a lie***

  – nouns that normally denote objects but that are assigned an eventive reading either through the process of type-coercion (Pustejovsky 1991), or through the processes of logical metonymy and coercion induced by temporal prepositions.

    (8) *I am finishing this* ***letter*** *rather hurriedly*

- pronouns related to previously mentioned events.

Differently from the Rich ERE and Event Nugget annotation, we do not annotate implied events indicated by nouns like *murderer* and *protestor* so as to make a clear distinction between events and entities and avoid confusion. Indeed, the annotation of implied events is a case of annotators' disagreement on event nugget tagging reported in Mitamura et al. (2015).

The factuality status of events does not impact the annotation: All events have to be annotated whether they are presented as a fact, a counterfact, or a possibility. This choice differs from what is done in the Light ERE annotation, in which only actual events are eligible to be annotated (Song et al. 2015).

3.2 Event Extent

Eventualities have different extents: The annotation of single-token, multi-token and discontinuous expressions is allowed as detailed subsequently. We decided to include continuous and discontinuous multi-token extents in the annotation so as to better capture together all the linguistic elements that are important components of meaning. This choice is in contrast with the minimal extent rule of TimeML and brings us closer to the Event Nugget annotation. However, we restrict the multi-token annotation to specific types of linguistic constructions (e.g., light and phrasal verbs) to reduce the risk of ambiguity. Indeed, the annotation of multi-token event nuggets is one of the main causes of disagreement because their annotation depends on the definitions of the different event types/subtypes (Mitamura et al. 2015).

Finite and non-finite verb forms: We annotate only the verbal head without auxiliaries of any form (multiple, modal, negative).

(9) *you **wish** to **know*** = 2 annotations

(10) *having been **destroyed** by the father*

Phrasal verb constructions: The main verb should be annotated together with the particle and/or the preposition forming the phrasal verb because they form a single semantic unit whose meaning cannot be understood by looking at the meaning of each single part. In case the verb and the preposition are separated, a discontinuous annotation should be performed.

(11) *I abjectly **stepped into** his cab* = 1 annotation

Light verbs: The whole predicate formed by the main verb and the following expression, usually a noun, is to be annotated even if not continuous.

(12) ***make** her **a visit*** = 1 annotation

(13) ***get a snap-shot*** = 1 annotation

Copular constructions: Past work (den Dikken and O'Neill 2016) distinguishes different types of copular constructions on the basis of a taxonomy of four copular elements: (i) support copula; (ii) predicational copula; (iii) equative copula; (iv) silent copula. The first two cases are to be annotated with a multi-token span including both the copula and the whole copula complement (14). As for equatives, whose linguistic status is unclear (Mikkelsen 2005), only the copula should be annotated (15).

(14) *Our welcome to Genoa **was not cheerful*** = 1 annotation

(15) *Dr. Jekyll **is** Mr. Hyde*

Periphrastic causative constructions: These are composed of a causative verb such as *cause*, *get*, *have*, or *make*, combined with another verb to express causation (Kemmer and Verhagen 1994). These two verbs should be annotated separately.

(16) *urging him to **make** his brother **drive** more carefully* = 2 annotations

Fixed expressions: Phrases, idioms, nominal expressions whose meaning cannot be understood from the individual meanings of their elements have to be annotated as a unique mention.

(17) *in order to* ***get rid of*** *him* = 1 annotation

(18) *a hostile* ***air raid*** *this evening* = 1 annotation

Nouns: Can be annotated within a multi-token or discontinuous expression if part of a copular construction, a light verb construction, or a fixed expression. In addition, named events such as "First World War" can also have a multi-token extent. In all the other cases, the noun itself should be annotated alone, without including determiners or adjectives.

(19) *both in* ***peace*** *and* ***war*** = 2 annotations

3.3 Semantic Classes

Each annotated event mention should be classified by assigning a value to the `CLASS` attribute. The classification we have designed is based on semantic criteria. We decided not to follow the TimeML classification because aspectual types and syntax do not have a primary importance for historians in the interpretation of texts. On the other hand, the types and subtypes defined in ACE, ERE, and Event Nugget are too limited and do not allow us to identify events comprehensively. On the contrary, we want our classification to be wide in terms of event type coverage: To this end, we re-elaborated the semantic categories of the Historical Thesaurus of the Oxford English Dictionary (HTOED) (Kay et al. 2009).

The HTOED has been defined over several decades with the aim of conceptualizing and classifying the meaning of the English language. The HTOED is extensively used to study English historical texts of different epochs (Roberts 2000; Alexander and Struan 2013) and a semantic tagger has been developed exploiting the information contained in the almost 800,000 entries of the thesaurus (Piao et al. 2017). However, this tagger does not use any machine learning approach but rather a look-up strategy combined with a set of algorithms for word sense disambiguation. In addition, the tool provides an all-words tagging without any specific focus on events.

The HTOED consists of a hierarchical structure made of a primary tripartite division (*External World*, *Mental World*, and *Social World*), 37 categories, and 377 sub-categories.[5] In HTOED a distinction is made between categories connected with a physical existence and those having a social dimension: Due to this subtle difference an event of movement can belong to the `TRAVEL AND TRAVELING`, the `SPACE`, or the `MOVEMENT` category. In other words, discerning between physical and social dimensions is ambiguous. Therefore, starting from the original complex and extremely fine-grained classification, we worked to find an appropriate level of granularity by merging categories with a common conceptual core.[6] This choice led us to create a unique class for events related to the

5 `http://historicalthesaurus.arts.gla.ac.uk/`.

6 WordNet supersenses (Ciaramita and Johnson 2003) have partial overlap with the HTOED and HISTO categories (see for example noun.possession/verb.possession in WordNet and the class POSSESSION in both HISTO and HTOED). However, there are several differences to highlight: First of all, supersenses are strongly based on PoS categories, since words that are semantically similar but grammatically different

concept of space (SPACE-MOVEMENT) and for those involving forces beyond scientific understanding or the laws of nature (RELIGION-SUPERNATURAL). In addition, we collapsed into the same class events in the area of production and trade of services and goods (ECONOMY), those in the public domain (LAW-AUTHORITY), and those involving all the types of living things and their health conditions (LIFE-HEALTH). Events connected to the faculties of the mind characterized by reasoning or knowledge are brought together in the MENTAL-ABSTRACT class, and instinctive or intuitive mental activities accompanied by a certain degree of pleasure or displeasure are joined in the EMOTIONS-EVALUATIONS class. Figure 1 provides a graphical representation of how our classes were defined starting from the HTOED categories.

3.3.1 Description of Semantic Classes. We now describe the classes, together with a set of examples. Event extension is highlighted in **bold**.

1. EARTH-ENVIRONMENT, eventualities related to geography (20), climate/weather conditions (21), environmental issues (22).

   (20) *the streets **are like caverns***

   (21) *It has been **raining** for days*

   (22) ***deforestation** has denuded the mountain-side*

2. LIFE-HEALTH, eventualities related to living things, namely, humans (23); animals and plants (24); including life, death, physical conditions, diseases, and medical treatments (25).

   (23) *he **was a Caprian paesant***

   (24) *oranges do not **grow up***

   (25) *in charges of contagious **diseases***

3. FOOD-FARMING, eventualities pertaining to food, food preparation and consumption (26); drink (27); agriculture and hunting (28).

   (26) *let us **breakfast** together*

   (27) *I luxuriously **sip** my coffee*

   (28) *an elderly man was **plowing** with a pair of oxen*

are labeled with different supersenses. Thus, the noun "trip" is annotated as noun.act whereas the verbs "to trip" and "to travel" are annotated as verb.motion. This is a main difference with respect to our classification. Besides, WordNet does not cover all the multi-token constructions taken into consideration in our work. Indeed, in WordNet there are entries for phrasal verbs (e.g., "to take away") and nominal expressions (e.g., "air raid") but not for light verbs. Moreover, adjectives are underrepresented in WordNet supersenses with only three possible classes: adj.ppl, adj.all, and adj.pert. On the contrary, adjectives have an important role in our annotation in the context of copular constructions. A last issue concerns the semantic coverage of supersenses, which do not account for classes that are important in our classification, such as RELIGION-SUPERNATURAL, AUTHORITY-LAW, HOSTILITY-MILITARY, and ECONOMY.

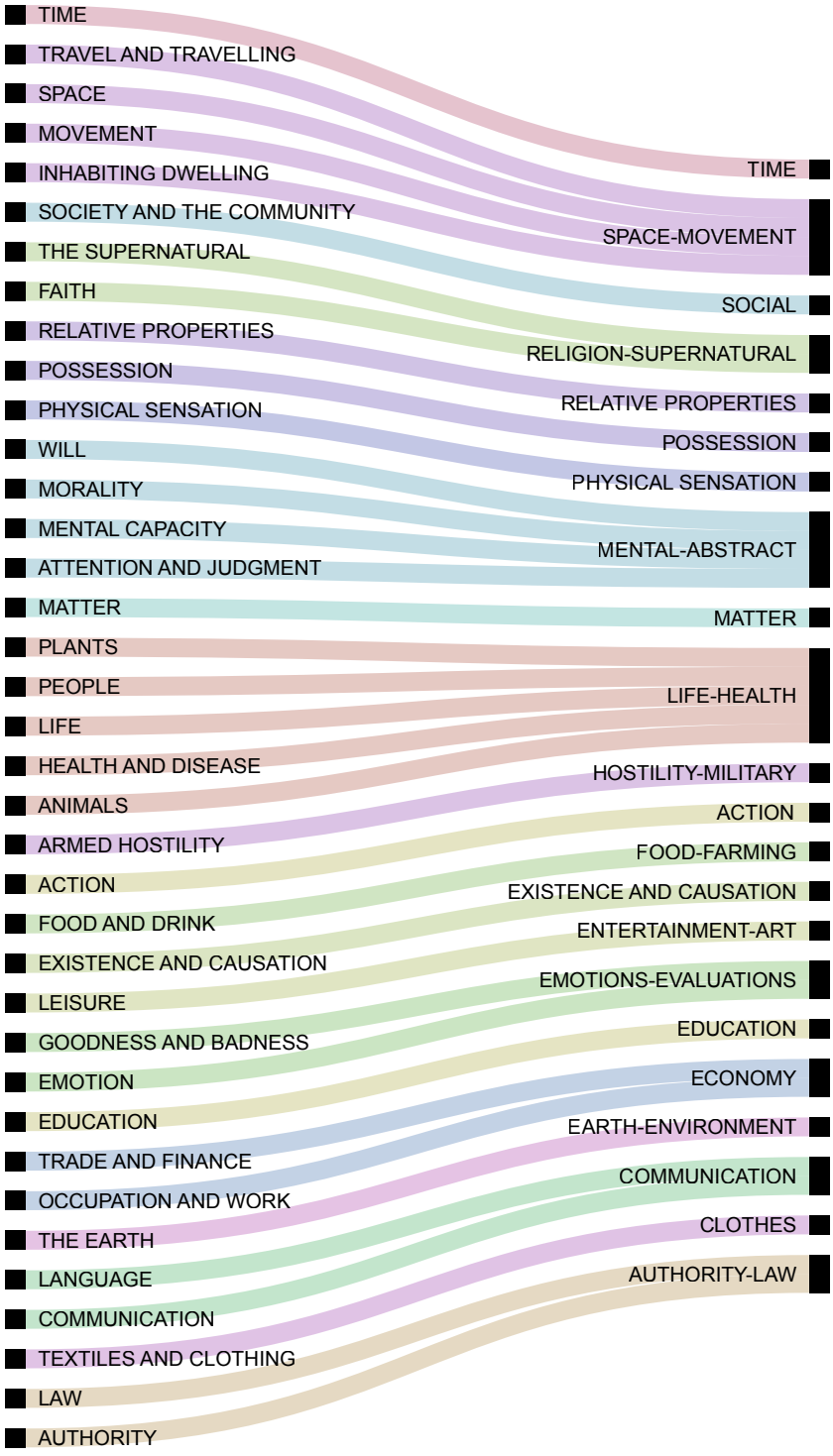

Figure 1
Mapping of the second-level HTOED categories (left) to our HISTO classes (left). Image created with RAWGraphs [Mauri et al. 2017].

4. CLOTHES, eventualities associated with textiles (29), clothes (30), and other personal belongings (31).

   (29) *they are renowned for their skill in* ***weaving***

   (30) *he* ***took off*** *his hat*

   (31) *it* ***is a rather heavy portmanteau***

5. MATTER, eventualities connected to substances and materials, their properties, constitution and conditions (32). This class includes terms relating to liquids (33), solids, gases, electricity, light (34), colors, and shapes.

   (32) *burdens that* ***seem too heavy***

   (33) *the miracle of* ***liquefaction***

   (34) *their full black hair* ***shines*** *like satin*

6. EXISTENCE-CAUSATION, eventualities relating to the concepts of being as in existential clauses (McNally 1998) (35), occurring (36), existing and causation (37), and their lack. It includes creation, destruction, damage, break, and demolition.

   (35) *in this court* ***are*** *a number of handsome sarcophagi*

   (36) *these* ***occurrences*** *are fanning a spirit of revenge*

   (37) *the cases* ***caused*** *me a genuine thrill*

7. SPACE-MOVEMENT, brings together all the eventualities pertaining to space (38), lack or end of movement (39), and travel 40.

   (38) *mitre of gold is* ***covered*** *with precious gold*

   (39) *the Temple of Minerva* ***standing*** *beside twelfth-century buildings*

   (40) *we* ***sailed*** *from New York six weeks ago*

8. TIME, eventualities associated with frequency (41), change, duration (42), age, and the spending of time (43).

   (41) *this* ***is the first time*** *the cup will leave France*

   (42) *the raid* ***lasted*** *for about half an hour*

   (43) *she* ***spent*** *some weeks at Sydney*

9. ACTION, general eventualities denoting not specific operations upon something like doing, using (44), trying, helping, finding, but also events and states relating to safety/danger (45), difficult/easiness, and success/failure.

(44) *the very best tobacco **used** in the cigar factories*

(45) *Doctor Antonio hesitated about **imperilling** her neck*

10. RELATIVE PROPERTIES, eventualities pertaining to measurements (46), numbers (47) (except those relative to the temporal dimension that have to be annotated with the TIME class), and quantities (48).

(46) *this island of Luzon **is so large***

(47) *because of a **reduction** in their wages*

(48) *these Paris delegates **are thirty-five***

11. RELIGION-SUPERNATURAL, eventualities related to religions (49), worship, and the supernatural (50).

(49) *the **high mass** is **celebrated***

(50) *the departed **haunt** the silent town*

12. MENTAL-ABSTRACT, includes all mental actions and processes (51), attention and judgment (52), and expressions of will (53).

(51) *Victor Emanuel **seems** to have **thought** that...*

(52) *he could **take care** of me and himself*

(53) *having **decided** to meet Zelphine and Angela*

13. EMOTIONS-EVALUATIONS, emotional actions, states and processes or eventualities expressing the lack of emotions (54, 55) (excitement/calmness, pleasure/suffering, compassion/indifference, courage/fear, love/hate).

(54) *the days brought me **enjoyment** and **delight***

(55) *the witnesses **were amazed** at the man's **calmness***

14. POSSESSION, includes eventualities associated with concepts such as having, not having, losing, taking, giving, allocating, acquiring, receiving, sharing (56, 57), and the opposition between wealthy and poverty (58).

(56) ***having** like it four colossal bronze lions at the base*

(57) *they would not **take** a large sum of money for the experience*

(58) *all his neighbors would testify to his **poverty***

15. COMMUNICATION, linguistic actions, states and processes connected to both the intellectual activity of speaking a language, naming things, producing speech acts

(59), and the social activity of expressing, transmitting, and receiving information in different ways through the media (60, 61).

(59) ***crying out**: "ecco, ecco, signora!"*

(60) *any one who can **write** letters as interesting as yours*

(61) *your mother will remember **reading** this story to me*

16. SOCIAL, eventualities involving the society in general or a specific community. The class includes social actions, states, and processes such as the participation, or lack of participation, in meetings (62) and relationships of different types: intimate, between family members, within groups (63) and associations.

(62) *the **meeting** was addressed by anarchists*

(63) *she will be calling me soon to **join** her*

17. HOSTILITY-MILITARY, eventualities related to different aspects of military life (64, 65) (operations, service, use of weapons), acts of hostility and peace.

(64) *they shall not **conscript***

(65) *America stands supremely for **peace***

18. AUTHORITY-LAW, eventualities associated with political and governmental activities (66) and in general to the exercise (67) or lack of authority (power, rule) but also to criminal activities and to the legal system (68) (legislation, legal power, punishments).

(66) *favorite candidate for the next municipal **election***

(67) *Commander Clifford **commanding** the Pampanga*

(68) *during the **trial** here in Buffalo*

19. EDUCATION, eventualities pertaining to teaching, learning, but also to the administration of educational institutions (69, 70).

(69) *he was **graduated** from Princeton University in 1906*

(70) *they are not **learning** anything*

20. ECONOMY, eventualities connected to money (71) (change of money, payments, and taxation), commerce (business affairs, trading operations, buying and selling), work and employment (72).

(71) *she **sold** her pearls to **raise money** to feed the poor*

(72) *Larcher **is a locksmith***

21. ENTERTAINMENT-ART eventualities related to entertainment (night-life, hobbies), arts (73) (performing art, music, visual arts), sports, and games in general (74).

    (73) *not content with this they **gave a dance** that same evening*

    (74) *the number of **games** to be **played** here will be at least three*

22. PHYSICAL SENSATION, eventualities related to the perception by senses (75) (touch, taste, smell, sight, hearing), but also sleeping/waking (76) and cleanness/dirtiness (77).

    (75) *the wind is scarce **felt**, though you may **hear** it **sighing***

    (76) *cancel all speaking engagements and **take a complete rest***

    (77) *the Neapolitan city **is even dirtier***

4. Data Set Construction

We applied the guidelines described in Section 3 to a newly created collection of historical texts named *Histo Corpus*. The following subsections describe the corpus and its annotation process with details on inter-annotator agreement.

4.1 Corpus Description

The *Histo Corpus* (henceforth, HC) consists of historical texts of two different genres, namely, travel narratives and news, published between the second half of the nineteenth century and the beginning of the twentieth century.

News have been taken from the newspaper portal Wikisource,[7] the Wikimedia Foundation Web site containing a digital library of source text transcriptions free of copyright. We selected news covering various topics, such as murders, conflicts, sports, movie reviews, obituaries, scientific discoveries, and gossip on celebrities. The historical nature of the texts and the diversity of topics covered by the news make them particularly interesting for annotation.

On the other hand, travel narratives are not much explored in computational linguistics. Exceptions are the ANC (American National Corpus) and GUM (Georgetown University Multilayer) corpora (Ide and Macleod 2001; Zeldes 2017), which, however, contain only contemporary texts, and the collection of historical German travel guides developed within the travel!digital project[8] (Czeitschner and Krautgartner 2017). Nevertheless, to the best of our knowledge, no corpus of travel narratives with event annotation has been released before.[9] We choose this particular genre because travel writings are powerful sources of information for many research areas, such as art history, ethnography, geography, and cultural history (Burke 1997). Being able to tag them automatically and to extract information about mentioned events would enable

7 https://en.wikisource.org/.

8 https://traveldigital.acdh.oeaw.ac.at/.

9 Travel guides in the travel!digital project are annotated following a domain-specific thesaurus that includes a very limited type of event, that is, tourist activities such as excursions and carriage rides.

Table 1
Statistics on the *Histo Corpus*.

	DOCS	TOKENS	PERIOD OF PUBLICATION
Travel Narratives	25	28,259	1865–1921
News	47	27,821	1883–1926
TOTAL	72	56,080	1865–1926

us, for example, to compare different sources and reconstruct the history of cultural sites, to study travelers' itineraries, or to analyze how the environment has been affected by specific event types.

Travel narratives included in *HC* are a subcorpus of a larger collection of texts we have created with the aim of fostering research on travel writings with digital and computational methods (Sprugnoli et al. 2017; Sprugnoli 2018b). More specifically, this collection consists of 57 books, for a total of 3,630,781 tokens: All the books are available in a cleaned text format and 30 of them are also distributed in TEI-XML on a dedicated Web site.[10] These books, both travel narratives (reports, diaries, collections of letters) and travel guides, were taken from Project Gutenberg,[11] are about Italy, were written by Anglo-American authors, and were published between the country's unification in 1861 and the beginning of the 1930s.

Table 1 shows details on the number of documents and tokens in *HC*, together with their period of publication. Even if HC is not as large as other corpora annotated with temporal information, at the moment of writing it is the largest available corpus annotated with events in the historical domain.

4.2 Corpus Annotation

The *Histo Corpus* was annotated following the guidelines described in Section 3 and using the Web-based CAT annotation tool (Bartalesi Lenzi, Moretti, and Sprugnoli 2012). This subsection contains description and results of the inter-annotator agreement performed to check the soundness of the guidelines and the feasibility of the proposed tasks. Then we give details on the annotated data with an analysis of the main differences between events annotated in the two genres forming the *Histo Corpus*.

4.2.1 Inter-Annotator Agreement. We measured the inter-annotator agreement (IAA) (Artstein and Poesio 2008) on a subset of the *Histo Corpus*, balanced between the two genres in terms of token number: one travel narrative and four news pieces about different topics (national and foreign policy, sport, scientific discoveries) were selected for a total of 1,200 tokens. Two annotators performed the work independently, using the guidelines reported in Section 3: One was one of the authors of the paper, and the other was not involved in the development of the guidelines. Both annotators have very good English proficiency and expertise in linguistic annotations.

Results of the IAA are reported here with different metrics. The Dice coefficient (Dice 1945) is given for the identification of event mentions, distinguishing between

10 https://sites.google.com/view/travelwritingsonitaly.
11 https://www.gutenberg.org/.

Table 2
Confusion matrix among IAA annotators for class annotation

	AUT	CLO	COM	EDU	EMO	ENT	ENV	EXI	REL	FOO	ACT	HOS	LIF	MAT	REL	MEN	PHY	POS	SOC	SPA	TIM	ECO
AUT	**10**	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0
CLO	0	**1**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
COM	0	0	**14**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
EDU	0	0	0	**2**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
EMO	0	0	0	0	**2**	0	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0
ENT	0	0	0	0	0	**3**	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0
ENV	0	0	0	0	0	0	**2**	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
EXI	0	0	0	0	1	0	1	**17**	0	0	0	0	0	0	0	0	0	0	0	2	0	0
REL	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0
FOO	0	0	0	0	0	0	0	0	0	**10**	0	0	0	0	0	0	0	0	0	0	0	0
ACT	0	0	0	0	0	1	0	2	0	0	**4**	0	0	0	0	0	0	0	0	0	0	0
HOS	0	0	0	0	0	0	0	0	0	0	0	**2**	0	0	0	0	0	0	0	1	0	0
LIF	0	0	0	0	0	0	0	0	0	0	0	0	**6**	0	0	0	0	0	0	0	0	0
MAT	0	0	0	0	0	0	0	1	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0
REL	0	0	0	0	0	0	0	0	0	0	1	0	0	0	**2**	0	0	0	0	0	0	0
MEN	1	0	0	0	2	0	0	5	0	0	1	0	0	0	0	**7**	0	1	0	0	0	0
PHY	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**10**	0	0	0	0	0
POS	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	1	0	**4**	0	0	0	0
SOC	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0
SPA	0	0	0	0	0	0	13	0	0	0	1	0	0	0	0	0	0	0	0	**31**	0	0
TIM	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	**3**	0
ECO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**10**

the agreement calculated on extensions perfectly detected by both annotators and the one measured on the number of annotated tokens shared by both annotators, thus considering also a partial match. In other words, with the Dice Coefficient we measure the agreement in determining whether each token is part or not of an event mention. For event classification, we calculated the Cohen's kappa (Cohen 1960) on mentions detected by both annotators, so as to also measure the pairwise agreement taking into consideration agreement that would be obtained by chance:

- EVENT MENTION DETECTION:
  - Dice Coefficient macro-average at tag level (perfect match): 0.85
  - Dice Coefficient macro-average at token level: 0.87
- EVENT CLASSIFICATION:
  - Cohen's kappa: 0.71

In linguistic annotation, a kappa score of 0.80 is considered the minimum threshold for data annotated with good reliability (Landis and Koch 1977; Carletta 1996). Our results on event mention detection are particularly good given the presence of multi-token and discontinuous mentions: The agreement on perfect match is only slightly lower than the one at token level (0.85 vs. 0.87), meaning that mentions can be detected in a consistent way. Disagreements were due to differences in the inclusion of prepositions in the event extent ("twister over") and to the non-identification of copular constructions ("the average speed was 44 miles per hour"). Another problematic case is given by polysemous event nominals like "story" in the following sentence, which may denote both an event and an information object: "a witness of the truth of the story."

As for event classification, results are lower but still satisfactory given the complexity of the task with 22 different options. Table 2 presents the results in a confusion matrix. Seven out of 22 classes achieved a perfect agreement (in brackets the number of occurrences in the annotated subset): `COMMUNICATION` (15), `EDUCATION` (2), `FOOD-FARMING` (1), `LIFE-HEALTH` (5), `PHYSICAL SENSATIONS` (11), `SOCIAL` (2), and `ECONOMY` (8). Disagreement in the other classes was registered, for example, for cases of figurative uses of verbs (e.g., "the white cap of Vesuvius worn generally like the caps of the Neapolitans"). Moreover, annotators tended to overuse the class `ACTION` as a backup category in case of uncertainty. The most frequent classes in the IAA subset are `SPACE-MOVEMENT` (44 occurrences) and `EXISTENCE-CAUSATION` (20 occurrences) with an agreement of 0.69 and 0.81, respectively.

By comparing the IAA on the *Histo Corpus* with the agreement reported for other schemes dealing with event annotation, it is worth noticing higher results both for the extent and the class of event mentions in our corpus. In TimeBank 1.2, the agreement is 0.81 on partial match, 0.71 on perfect match, and 0.67 on class assignment.[12] For the data used in the Event Nugget task in 2015, the agreement on event detection does not reach 0.80 and is below 0.70 on event classification (Song et al. 2016).

4.2.2 Annotated Data. Table 3 reports the number of annotated events in HC per class and text genre. News and travel narratives show, for almost all the event classes, a

12 Data reported in the TimeBank 1.2 documentation: `http://www.timeml.org/timebank/documentation-1.2.html`.

Table 3
Annotated events per class and text genre together with the total amount of annotations. The asterisk indicates whether the class has a statistically significant difference in the distribution over the two genres.

CLASS	NEWS	TRAVEL	TOTAL
SPACE-MOVEMENT*	791	963	1,754
COMMUNICATION*	571	377	948
ACTION*	516	315	831
MENTAL-ABSTRACT	420	419	839
EMOTIONS-EVALUATIONS*	239	450	689
EXISTENCE-CAUSATION*	360	296	656
PHYSICAL SENSATIONS*	200	324	524
LIFE-HEALTH*	215	144	359
POSSESSION	173	166	339
HOSTILITY-MILITARY*	260	25	285
TIME	119	120	239
AUTHORITY-LAW*	205	9	214
ENTERTAINMENT-ART*	103	68	171
ECONOMY*	115	46	161
RELATIVE PROPERTIES	67	67	134
SOCIAL*	96	32	128
MATTER*	37	86	123
ENVIRONMENT*	23	71	94
FOOD-FARMING*	13	56	69
CLOTHES	37	21	58
RELIGION-SUPERNATURAL*	2	27	29
EDUCATION	16	7	23
TOTAL*	**4,578**	**4,089**	**8,667**

statistically significant difference (at $p < 0.05$ and calculated with the z test[13]) in their distribution.

The high occurrence of events belonging to the SPACE-MOVEMENT class in both genres is due to the broad definition of the class that covers the three main concepts of motion (Sablayrolles 1995)—namely, locations, positions, and postures—and both factive and fictive motions. Examples of change of location (78), position (79), and posture 80 are given here. These are cases of factive motions, whereas Example (81) contains events of fictive motions, that is, "linguistic instances that depict motion with no physical occurrence" (Talmy 1996):

(78) *Marcel Renault **arrived** first*

(79) *the pigs used to **run about** in the principal streets of Naples*

(80) *a man **lay** in one of the entrances to the Union Station*

(81) *a deep ravine **surrounded** by mountains*

13 The z test is a parametric statistical test used to verify if the mean value of a distribution differs significantly from a certain reference value (Sprinthall 2003).

Additionally, the range of COMMUNICATION events is wide and particularly relevant in the news that typically reports the testimonies of observers and witnesses of what is recounted (see Examples (82) and 83):

(82) *he **told** Inspector Fairey*

(83) *he **admitted** that the council may have made mistakes*

The predominance of LIFE-HEALTH, HOSTILITY-MILITARY, and AUTHORITY-LAW is characteristic of news only: These classes cover events expressing, among others, murders and injuries, local riots, war offences, public administration, and judicial process, therefore they are particularly frequent in news about crimes, conflicts, and politics.

On the contrary, the PHYSICAL SENSATIONS class is strongly represented in travel narratives, in which the writer reports their experiences with local people and local environments.

As for the extent, 897 event mentions in the news and 860 in travel narratives are annotated with a multi-token span: These numbers correspond to the 19.6% and the 21.4% of the total number of events in the two genres, respectively. The majority of multi-token events are copular constructions with the verb "to be" (45.5%) but other verbs used as copulae are present as well, for example "to become" and "to feel." These constructions are mainly annotated with the class EMOTIONS-EVALUATIONS, while the second most common class for multi-token event mentions is SPACE-MOVEMENT. This class covers many phrasal verbs, such as "go out" and "go away."

The last row of Table 3 shows that the difference in the total number of annotated events in news and travel narratives is also statistically significant, with the former having a higher occurrence of event mentions (4,578 vs. 4,089).

5. Events in Historical Texts: Automatic Annotation

After having defined annotation guidelines and manually tagged a corpus accordingly, as described in the previous section, we report here on experiments on the automatic detection and classification of event mentions.

Experiments were carried out using the annotated *Histo Corpus*, divided into a training, test, and dev set (Section 5.1). For classification, we followed two different approaches. On the one hand, we implemented two CRF classifiers detailed in Section 5.2: One is aimed at identifying the correct span of event mentions and the other at assigning the correct class to each event mention. This latter task implies also the identification of mentions from raw text. In other words, no golden event mentions are given in input to the system. For the CRF classifiers we provide an analysis of features and of the impact of different context windows on the precision, recall, and F1-score. On the other hand, we used a BiLSTM implementation for sequence tagging: Also in this case, both tasks (event detection and event classification) were taken into account. This implementation relies on the findings in Reimers and Gurevych (2017a) and does not require any feature engineering: It is based on a neural architecture and on the use of dense vectors representing words. In Section 5.3 we describe the general architecture of the system and the results obtained by evaluating different hyperparameters' options and pre-trained word embeddings.

```
<Document doc_name="file.txt">
<token t_id="1" sentence="0" number="0">We</token>
<token t_id="2" sentence="0" number="1">set</token>
<token t_id="3" sentence="0" number="2">forth</token>
<token t_id="4" sentence="0" number="3">at</token>
<token t_id="5" sentence="0" number="4">eight</token>
<token t_id="6" sentence="0" number="5">o'clock</token>
<token t_id="7" sentence="0" number="6">.</token>

<Markables>
<EVENT_MENTION m_id="1" comment="" class="SPACE-MOVEMENT">
<token_anchor t_id="2"/>
<token_anchor t_id="3"/>
</EVENT_MENTION>

</Markables>

<Relations>
</Relations>
</Document>
```

```
MENTION DETECTION ONLY TASK
We          O
set         B-EVENT_MENTION
forth       I-EVENT_MENTION
at          O
eight       O
o'clock     O
.


DETECTION+CLASSIFICATION TASK
We          O
set         B-SPACE_MOVEMENT
forth       I-SPACE_MOVEMENT
at          O
eight       O
o'clock     O
.           O
```

Figure 2
Example of a file in the CAT XML format (left) and in the corresponding converted BIO/IOB2 notation (right) for the two tasks.

5.1 Data Preparation

As a first step, we automatically converted annotated files from the CAT format to the BIO/IOB2 notation. The former is the stand-off XML format of the CAT annotation tool: In it, different annotation layers are contained in separate document sections and related to each other and to the source text through pointers. The latter is a tagging scheme in which a "B-" tag marks the first token of an annotated segment (in our case a segment is an event mention), "I-" is used for all the other tokens within the span of the same segment, and "O-" marks tokens that do not belong to the segment (Sang and Veenstra 1999). We chose the BIO/IOB2 notation because for the BiLSTM architecture it has proven to perform better than other notations such as IOB1 (Reimers and Gurevych 2017a), in which the "B-" tag marks the beginning of an annotated segment only when it immediately follows another annotated segment. Following the recommendation of Reimers and Gurevych (2017a), we preferred the BIO/IOB2 scheme to the IOBES one as well, because the latter tends to generate a bigger overhead.

Figure 2 shows an example of CAT and BIO/IOB2 formats. For the mention detection task, the "B-EVENT_MENTION" and "I-EVENT_MENTION" tags are used to indicate the span of each event; for the classification task, tags are used to specify the event class and, implicitly, its extension.

After the conversion, we divided the *Histo Corpus* into training (80% of the whole corpus), test (10%), and development sets (10%). The files were chosen randomly as for class value but we balanced the distribution in each section across the two genres.

5.2 CRF Classifiers

For the first set of experiments, we implemented two linear CRF classifiers using CRFsuite, a software for labeling sequential data: It contains different state-of-the-art training methods and an integrated evaluation functionality to compute Precision, Recall, and F1-score on test data.[14] In all the experiments, we used the default training

14 http://www.chokkan.org/software/crfsuite/.

algorithm of CRFSuite (Limited-memory Broyden-Fletcher-Goldfarb-Shanno) with L1 regularization. In addition, we put a threshold to ignore features whose frequency of occurrence in the training data is less than 2 and made CRFsuite generate both state and transition features.

As for features, we chose a simple set of three beyond the token itself: (i) lemma, (ii) PoS, and (iii) text genre. The first two were extracted by processing the texts in the *Histo Corpus* with Stanford CoreNLP (Manning et al. 2014) and the third marks the opposition between news and travel narratives at document level. Although other, more semantically rich features could be used, we limit our feature set to a few basic ones, in line with the setting adopted with the BiLSTM approach (Section 5.3), which relies only on word embeddings. In the next subsection we present the results of several experiments carried out on the development set: In particular, we analyze the impact of the features and of the size of the context window on the performance of the classifiers.

5.2.1 Feature Selection. We adopted a backward selection approach: We trained and tested the model with all the features and then removed them one by one to identify the best feature set. The model was obtained on the training set and then tested on the dev set. The results of this feature selection are reported in Table 4 for both the tasks of event mention detection and of event classification starting from raw text, thus including the identification of mentions. We provide information about the macro-average precision (P), recall (R), and F1 calculated considering a context window of $[+/-2]$ for all features. The best combination of features is given in bold.

As for the task of event mention detection, all the combinations of features beat the baseline. More specifically, without information on lemma precision improves (+1.32 points) but the overall F1 slightly drops (-0.14). PoS proved to be the most important feature: without this grammatical information, all the evaluation measures significantly drop (-1.59 for precision, -6.24 for recall, and -4.5 for F1 with respect to the configuration with all the features). Text genre improves neither recall nor F1. The feature combination including lemma and PoS shows an improvement over the baseline, especially in terms of recall (+9.07).

As regards event classification with no golden mentions, the results are low for all the configurations, with an F1 below 30%. Moreover, precision and recall are not balanced, with a difference ranging between 13.81 and 15.85 points, depending on the

Table 4
Performance, in terms of precision (P), recall (R), and F1, of the CRF model on the development set for both event mention detection and event detection+classification tasks with different settings of features.

	MENTION DETECTION ONLY			DETECTION+ CLASSIFICATION		
	P	**R**	**F1**	**P**	**R**	**F1**
ALL FEATURES	86.60%	80.56%	83.33%	30.05%	25.24%	28.45%
- without lemma	**87.92%**	79.48%	83.19%	33.43%	19.19%	22.19%
- without PoS	85.01%	74.32%	78.83%	38.83%	22.98%	27.26%
- without genre	86.74%	**80.60%**	**83.41%**	**41.00%**	**25.46%**	**29.13%**
BASELINE (only tokens)	82.72%	71.53%	76.15%	36.29%	20.15%	24.56%

Table 5
Performance of the CRF classifier on the development set on event mention detection and on event mention detection+classification with different context windows.

	MENTION DETECTION ONLY			DETECTION+ CLASSIFICATION		
CONTEXT	**P**	**R**	**F1**	**P**	**R**	**F1**
0	81.41%	70.06%	74.10%	38.29%	23.99%	27.73%
±1	85.42%	79.04%	81.95%	37.72%	**25.99%**	**29.27%**
±2	86.74%	**80.60%**	**83.41%**	**41.00%**	25.46%	29.13%
±3	**87.89%**	79.72%	83.34%	35.79%	24.21%	27.65%
±4	87.06%	79.49%	82.87%	32.37%	23.28%	25.78%

feature. This difference is even more evident in the baseline (16.14 points). Information about the lemma and PoS of each token increases both precision and recall. Information on genre is not helpful and the best combination of features includes only lemma and PoS with an improvement over the baseline in terms of precision (+4.71), recall (+5.31), and F1 (+4.57).

5.2.2 Impact of Context Size. A second aspect to evaluate is the size of the context window around the token to be classified. To this end, we tested whether the choice of having a context window of [±2] positions is optimal. Table 5 shows the performance of the CRF classifier for event mention detection and for event detection+classification trained with the best feature selection (token + lemma + PoS) considering different context windows: no context window (0), [±1], [±2], [±3], [±4]. For each option we give the value of the macro-average precision, recall, and F1.

In the detection of mention extent, recall proves to be very sensitive to context window: By using single token features only (i.e., by considering a context window equal to 0) precision is already above 80%, whereas recall is 70.06%. When using a window of [±1], precision increases (+4.01) but recall shows an evident boost (+8.98). The best performance is achieved with a context of [±2], which also provides more balanced results between precision and recall (6.14 points).

For the other task, precision fluctuates considerably by changing the window, with a difference between 0.57 and 5.21 points, depending on the number of tokens in the context. The best F1 (29.27%) is given by a context of [±1]. However, as already noted with the experiments on features, precision and recall are not balanced, showing a difference of 11.73 points.

5.3 BiLSTM Approach

Our second approach is based on the use of an implementation of BiLSTM developed at the Ubiquitous Knowledge Processing Lab, Technische Universität Darmstadt.[15] We chose this implementation because the authors provide several tests with various hyperparaments that we took as a reference for our own tests. They also suggest a default

15 https://github.com/UKPLab/emnlp2017-bilstm-cnn-crf.

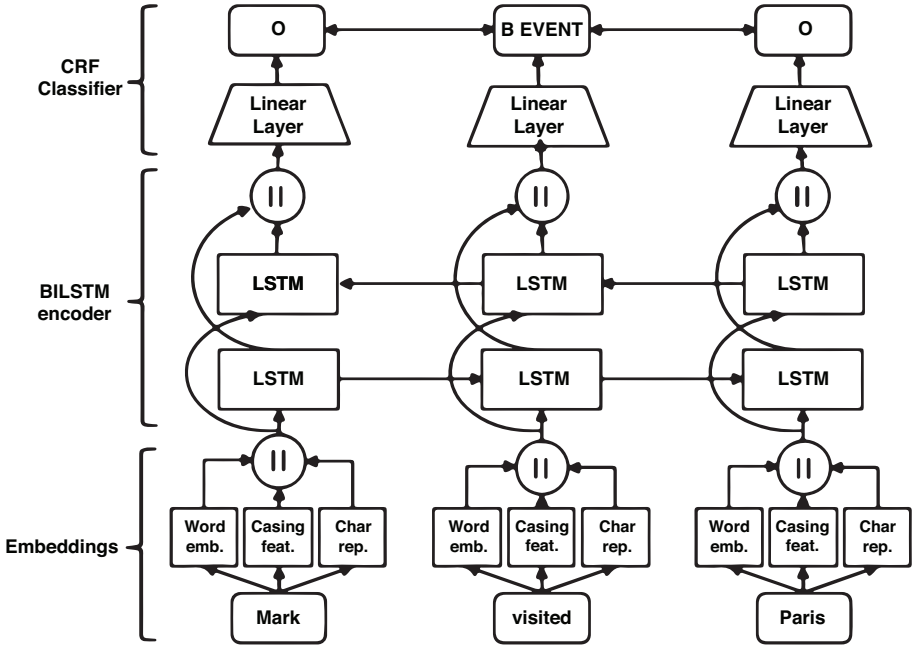

Figure 3
Architecture of the BiLSTM network with a CRF-classifier adapted from Reimers and Gurevych (2017a).

configuration for various NLP tasks, among which is the identification of events following TimeML annotation that we adopted as our starting point for our own experiments (Reimers and Gurevych 2017a). Figure 3, adapted from Reimers and Gurevych (2017a), displays the main architecture of the system with a CRF classifier as the final layer of the network—that is, with the best configuration we tested. Each word is mapped to a pre-trained word embedding and analyzed to detect its casing (i.e., numeric, mainly numeric, lower case, or upper case), and each character of the word is mapped to the corresponding character-level representation vector. Information about word embeddings, casing, and character embeddings is concatenated to be fed to the BiLSTM encoder. After the network has run from the beginning to the end of the sentence and vice versa, its output vectors are concatenated and fed to the last layer that can be a CRF classifier (as shown in Figure 3) or a Softmax classifier. This second option was tested as well, together with other hyperparameters, and the results are reported in the following subsection. This architecture does not require feature engineering, but only pre-trained word embeddings and a corpus of labeled data.

5.3.1 Testing Different Hyperparameters. This subsection reports on the performances obtained on the two tasks—event mention detection only and event detection+ classification—using the BiLSTM implementation previously described. Results are in terms of precision, recall, and F1: Given that different seed values can produce very different results (Reimers and Gurevych 2017b), we run the system three times, take the test score from the epoch with the highest results on the development set, and then we compute the average score.

Table 6
Average precision (P), recall (R), and F1 over three runs of the BiLSTM system with the configuration suggested by Reimers and Gurevych (2017a).

TASK	P	R	F1
MENTION DETECTION ONLY	82.50%	83.53%	82.99%
DETECTION+CLASSIFICATION	63.46%	62.93%	63.19%

As for the experimental settings, we took as a reference the setup suggested in Reimers and Gurevych (2017a), summarized here:[16]

- Mini-batch size: 8
- Recurrent units: 100
- Number of LSTM layers: 2
- Dropout: variational [0.25, 0.25]
- Classifiers: CRF
- Optimizer: nadam (Adam with Nesterov momentum) (Dozat 2016)
- Character representation: Convolutional neural networks
- Word embeddings: Komninos and Manandhar (2016)

Starting from this configuration, we performed a set of experiments changing several hyperparameters in order to identify the best options for our tasks. Below we report the results of these experiments to be compared to the ones in Table 6, which were obtained using the previously listed configuration.

All the experiments whose results are reported in the remainder of this subsection have been carried out with an early stopping after 10 epochs if the score on the development set did not increase. The implementation is based on Keras 1,[17] with Theano 1.0.0[18] as backend. In the remainder of this subsection we report on the different configurations we tested: six optimization algorithms, two character representations, two classifiers, and nine pre-trained word embeddings.

Optimizer. Optimization algorithms are used to update the model parameters with the aim of reducing a cost function. Over the years, different algorithms have been proposed: for example, SGD (Robbins and Monro 1951), Adam (Kingma and Ba 2014), Nadam (Dozat 2016), Adagrad (Duchi, Hazan, and Singer 2011), Adadelta (Zeiler 2012), and RMSProp (Tieleman and Hinton 2012). Table 7 gives details on the performance of these optimizers on event mention detection and on event classification with no golden mentions.

In both tasks, the worst results are achieved with SGD. The difference with respect to the other optimizers is evident in particular in the classification task, where SGD obtained an F1 of only 48.69%, whereas all the other algorithms have an F1 above 62%. This is unsurprising, because SGD is an optimization algorithm known to require

16 In their paper, Reimers and Gurevych (2017a) take into consideration various NLP tasks, including event detection in accordance with the TimeML guidelines, thus considering only single-token mentions. For this task the authors test numerous configurations highlighting the hyperparameters that have a high impact on the performance in terms of F1. Our tasks are, however, more complex, given that they include the identification of multi-token and discontinuous mentions and their classification.

17 `https://keras.io/`.

18 `http://deeplearning.net/software/theano/`.

Table 7
Results of the BiLSTM system with different optimization algorithms on the event mention detection only task and on the event detection+classification task.

	MENTION DETECTION ONLY			DETECTION+ CLASSIFICATION		
OPTIMIZER	**P**	**R**	**F1**	**P**	**R**	**F1**
Nadam	**82.5%**	83.53%	**82.99%**	63.46%	62.93%	63.19%
Adam	80.37%	83.47%	82.99%	62.7%	**63.90%**	63.27%
SGD	79.73%	80.94%	80.51%	51.50%	46.20%	48.69%
Adagrad	81.50%	83.30%	82.40%	62.8%	62.43%	62.61%
Adadelta	80.15%	**84.20%**	82.14%	62.97%	63.13%	63.05%
RMSProp	80.90%	83.03%	81.91%	**63.93%**	62.70%	**63.32%**

careful tuning of its learning rate, which we do not perform here, and also yields worse results in the beginning of training but achieves better generalization later (Keskar and Socher 2017). Besides, with a 95% confidence interval, we observe that only the difference between Nadam and SGD is statistically significant in the mention detection task, while in the other task all optimizers yield a performance that is statistically significantly better than SGD (but with no significant difference among them).

Character Embeddings. The architecture we adopted implements two different approaches to derive character representations: One is based on a convolutional neural network (CNN) that takes into account only character trigrams without considering their position inside the word (Ma and Hovy 2016); the other uses a BiLSTM network considering all the characters of the word and also their position, thus distinguishing between characters at the beginning, in the middle, and at the end (Lample et al. 2016). Table 8 shows that by using this latter approach on the mention detection task, F1 is higher thanks to an increase in recall (+1.7 with respect to the CNN approach). This result confirms the findings of Reimers and Gurevych (2017a)—namely, that LSTM character embeddings are the best performing in the TimeML event detection task.

As for event classification, the right hand-side of Table 8 shows that the two character-based representations do not contribute much to the overall performance: the difference between them is minimal with a variation of only a few decimals. In the experiments reported in Reimers and Gurevych (2017a), discarding character embeddings is never the best option. However, in our case the highest precision, recall, and F1 are achieved without using them.

Table 8
Performance with different character embeddings options on the event mention detection only task and on the mention detection+classification task.

	MENTION DETECTION ONLY			DETECTION+ CLASSIFICATION		
	P	**R**	**F1**	**P**	**R**	**F1**
CNN	82.50%	83.53%	82.99%	63.46%	62.93%	63.19%
LSTM	81.40%	**85.23%**	**83.37%**	63.86%	63.30%	63.57%
NONE	**82.53%**	83.65%	83.04%	**63.93%**	**63.70%**	**63.81%**

Table 9
Precision, Recall, and F1 score with the CRF and Softmax classifiers in the event mention detection only task and in the event detection+classification task.

	MENTION DETECTION ONLY			DETECTION+ CLASSIFICATION		
	P	**R**	**F1**	**P**	**R**	**F1**
CRF	**82.50%**	**83.53%**	**82.99%**	**63.46%**	**62.93%**	**63.19%**
Softmax	81.10%	82.30%	81.69%	62.67%	62.57%	62.61%

Classifier. The last layer of the network can be configured as a CRF or a Softmax classifier. The main difference between the two classifiers is that in Softmax each token is seen as isolated, without considering dependencies between the tags in a sentence, whereas in CRF correlations between tags are taken into account. Our results are in contrast to the ones discussed in Reimers and Gurevych (2017a): Softmax performs better in the TimeML event detection task because only single-token events are annotated, thus no information about tag dependencies is needed. As reported in Table 9, instead, using a CRF classifier as the last layer achieves better results for all three evaluation metrics in both tasks.

Pre-trained Embeddings. In recent years, pre-trained word vectors have become important resources largely adopted to deal with many NLP tasks (Collobert et al. 2011) and many pre-trained word embeddings have been released. Beyond Komninos and Manandhar embeddings (*Komn*),[19] we tested other resources available online, namely:

- GloVe, with both 300 and 100 dimensions (*GloVe300 - GloVe100*)[20] (Pennington, Socher, and Manning 2014), trained on a corpus of 6 billion tokens consisting of the 2014 English Wikipedia and Gigaword 5;
- *GoogleNews*, with 300 dimensions and trained on a subset of the Google News corpus (about 100 billion words)[21] (Mikolov et al. 2013);
- Levy and Goldberg embeddings (*Levy*),[22] with 300 dimensions and produced from the English Wikipedia on the basis of dependency-based contexts (Levy and Goldberg 2014);
- *fastText*, with 300 dimensions and trained on the English Wikipedia using character *n*-grams [23] (Bojanowski et al. 2017).

By taking into consideration the previously listed pre-trained embeddings, we cover different types of word representation: GloVe and GoogleNews are based on linear bag-of-words contexts, Levy and Komn on dependency parse-trees, and fastText on a bag of character *n*-grams. We also created additional historical word embeddings by processing a subset of the Corpus of Historical American English (COHA) (Davies 2012) with GloVe, fastText, and Levy and Goldberg's code. The subset of COHA we have chosen contains 36,856 texts published between 1860 and 1939 for a total of more than

19 https://www.cs.york.ac.uk/nlp/extvec/.
20 https://github.com/stanfordnlp/GloVe
21 https://code.google.com/archive/p/word2vec/.
22 https://levyomer.wordpress.com/2014/04/25/dependency-based-word-embeddings/.
23 https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md.

Table 10
Results obtained with different pre-trained word embeddings for the event mention detection only task and for the event detection+classification task.

	MENTION DETECTION ONLY			DETECTION+ CLASSIFICATION		
EMBEDDINGS	**P**	**R**	**F1**	**P**	**R**	**F1**
Komninos	**82.50%**	**83.53%**	**82.99%**	**63.46%**	**62.93%**	**63.19%**
FastText	79.47%	82.00%	81.25%	62.18%	61.79%	62.02%
GoogleNews	80.60%	81.70%	81.15%	63.20%	62.67%	62.93%
GloVe300	80.07%	79.73%	79.89%	61.87%	59.57%	60.69%
GloVe100	79.30%	80.90%	80.13%	60.23%	58.10%	59.16%
Levy	79.40%	83.13%	81.21%	62.10%	60.83%	61.44%
HistWords Google Aver.	78.93%	77.97%	78.43%	61.03%	57.80%	59.36%
HistWords Google Concat.	79.30%	78.57%	79.03%	62.47%	59.37%	60.87%
HistWords COHA Aver.	77.07%	77.93%	77.48%	56.60%	52.23%	54.34%
HistWords COHA Concat.	79.04%	79.34%	79.20%	59.60%	55.13%	57.29%
HistoLevy	80.47%	83.47%	81.95%	63.20%	61.70%	62.46%
HistoFast	79.90%	81.00%	80.44%	59.80%	55.93%	57.78%
HistoGloVe	80.47%	80.24%	80.51%	62.73%	60.17%	61.42%

198 million words. Texts belong to four main genres (fiction, newspaper, magazine, non-fiction) balanced within each decade. The word embeddings thus trained (*HistoGlove*, *HistoFast*, and *HistoLevy*) have 300 dimensions and are publicly available online.[24]. We also experimented with HistWords,[25] a collection of pre-trained word2vec historical word vectors divided by decades (Hamilton, Leskovec, and Jurafsky 2016). We have extracted the embeddings of the decades between 1860 and 1930, constructed via both Google N-grams and the COHA corpus, and then combined them using two strategies: (i) making the average between vectors of the different decades; and (ii) concatenating the vectors. The first strategy originated embedding dimensions of size 300 (*HistWords Google Aver.* and *HistWords COHA Aver.*), whereas the second strategy produced embeddings with 2,400 dimensions (*HistWords Google Concat.* and *HistWords COHA Concat.*).

Table 10 contains results obtained with the tested word embeddings for event detection and event classification. In both tasks, the Komninos and Manandhar embeddings perform best: the configuration including them is the only one that reaches almost 83% F1 for event detection and exceeds 63% for event classification. This means that capturing both semantic and syntactic similarities between words is crucial for the tasks. Also, Levy and Goldberg embeddings are dependency-based, but their precision falls below 80% in event detection. The main difference between the two representations is that Komninos and Manandhar extended the skipgram model including more types of co-occurrences within the dependency graph, thus they better capture the functional properties of words. As for GloVe, there is not much difference between the two dimensions (300 and 100). However the overall results are almost 3 points lower for event detection and 4 points lower for classification compared to the model employing

24 Our historical embeddings are available on GitHub: `https://github.com/dhfbk/Histo` Original raw texts extracted from COHA cannot be distributed because of copyright restrictions: `https://www.corpusdata.org/restrictions.asp`.

25 `https://github.com/williamleif/histwords`.

the Komninos and Manandhar embeddings. No improvement is registered for either GoogleNews or fastText but the former performs better in the detection+classification task than in the mention detection only task.

As for historical embeddings, concatenating HistWords embeddings yields better results than averaging them, even if neither of the two strategies reaches 80% F1 on the mention detection task and only the *HistWords Google Concat*. exceeds 60% on the detection+classification task.[26] For what concerns the embeddings we trained, the contribution of *HistoGloVe* and *HistoFast* is not helpful. However, they both achieve higher results in terms of F1 than with Glove in the mention detection only task. *HistoGlove* performs better than Glove also in the detection+classification task. Results obtained with *HistoLevy* are very promising: It achieves the second best F1 score (81.95%) in the mention detection task and the third best score in the classification task (62.46%), with only a modest difference with respect to GoogleNews in terms of F1 (0.47). These scores confirm that dependency-based embeddings have a positive impact on our tasks.

It is important to note that the amount of training data strongly affects the quality of word vectors, because more data produce more accurate vectors (Mikolov et al. 2013). However, our historical word representations were trained on a corpus that is much smaller than the corpora used to build the other embeddings (for example, *GoogleNews* embeddings are trained on about 100 billion words, whereas the COHA subset consists of just 198.7 million words). This might be the reason why we achieved a lower performance.

6. System Comparison and Discussion

The evaluations described in the previous sections allowed us to identify the best configurations for our tasks and for the two approaches under investigation, that is, CRF and BiLSTM.

For the task of mention detection, the best CRF classifier we release is based on a combination of three features (token, lemma, and PoS) and a context window of [±2]. For the same task, we set the neural architecture with the following parameters:

- Mini-batch size: 8
- Recurrent units: 100
- Number of LSTM layers: 2
- Dropout: variational [0.25, 0.25]
- Classifiers: CRF
- Optimizer: Nadam
- Character representation: LSTM
- Word embeddings: Komninos and Manandhar (2016)

The left-hand side of Table 11 reports the performance of the best models we obtained for the detection of event mentions evaluated on the test set. We also compute a baseline (i.e., a CRF classifier trained having only tokens as features). The difference between the CRF classifier and the BiLSMT model in terms of F1 is minimal (0.05). However, it is interesting to notice that the former has a higher precision whereas the second has a higher recall. This means that the neural architecture is more able to generalize the observations of events from the training data.

26 We also tested the impact of embeddings built on documents from single decades on the tasks but results are not better: 57.50% F1 on mention detection and 77.94% F1 on event detection and classification.

Table 11
Results of the CRF classifier and the BiLSTM model with the best configuration for the event mention detection only task and for the detection+classification task.

	MENTION DETECTION ONLY			DETECTION+ CLASSIFICATION		
	P	**R**	**F1**	**P**	**R**	**F1**
CFR	84.95%	82.36%	83.57%	37.21%	27.65%	29.09%
BiLSTM	82.30%	85.00%	83.62%	66.20%	62.70%	64.39%
Baseline	80.35%	74.64%	77.14%	31.25%	19.26%	21.33%

The task dealing with both event detection and classification needed different configurations. The CRF classifier was trained with tokens, PoS, and lemmas, as in the other task, but with a context size window of [±1]. In the BiLSTM network two different hyperparameters turned out to achieve better performance with respect to the ones adopted for the mention detection only task. More specifically, we applied the RMSprop optimizer, instead of Nadam, and we did not use any character-based representation. Scores for this task are reported on the right-hand side of Table 11 and compared to the baseline obtained, also in this case, by training a CRF classifier only with tokens as features.

The BiLSTM network performs remarkably better than CRF with a difference of more than 28.99 points in terms of precision, 35.05 points in terms of recall, and 35.30 as for F1. This shows that, whereas for the detection of the mentions, the lack of semantic information as a feature of the CRF classifier has no relevant impact; this plays an important role in the mention classification. Also, the BiLSTM network does not rely on complex linguistically informed features, but the embeddings alone are enough to capture well the meaning and the context of mentions, which are necessary to assign the correct class. Overall, both approaches are skewed toward precision: This bias is more evident in the CRF, whereas the BiLSTM network achieves more balanced results. A detailed comparison of the scores at the level of event classes is given in Figure 4.

Both approaches failed to classify events of the classes `FOOD-FARMING` and `FAITH`, which had very few occurrences in both the training and the test set. The BiLSTM model wrongly classified `EDUCATION` events: In particular, it assigned the class `MENTAL-ABSTRACT` to the verb "to learn." This annotation is not totally incorrect from the semantic point of view, given that learning is a mental process. On the other hand, the CRF classifier did not assign the correct value to any of the events in the `HOSTILITY-MILITARY` and `ENVIRONMENT` class. As for the latter, the performance is not high with the BiLSTM model (F1=31.58%) either because it failed in the classification of nominal events (e.g., "storm", "tempest") and properly classified the verb "to fall" only when the subject, belonging to the environmental domain, was close to the verb. In the following sentence, for example, "falling" is annotated with the right class whereas "fell" is annotated with the class `SPACE-MOVEMENT`: *rain commenced falling at 8:10 p.m. , and between 8:14 and 8:26 one-fifth of an inch fell*.

The different performance between the two approaches is very evident for some classes: As for `AUTHORITY-LAW`, BiLSTM is able to recognize verbs, nouns, and expressions related to judicial processes and government-sanctioned practices that CRF do not even annotate as events (e.g., "to be sentenced," "confinement," "to be charged"). CRF also fails in the recognition of some aspectual events that BiLSTM correctly annotates with the class `TIME` (e.g., "to cease," "to commence").

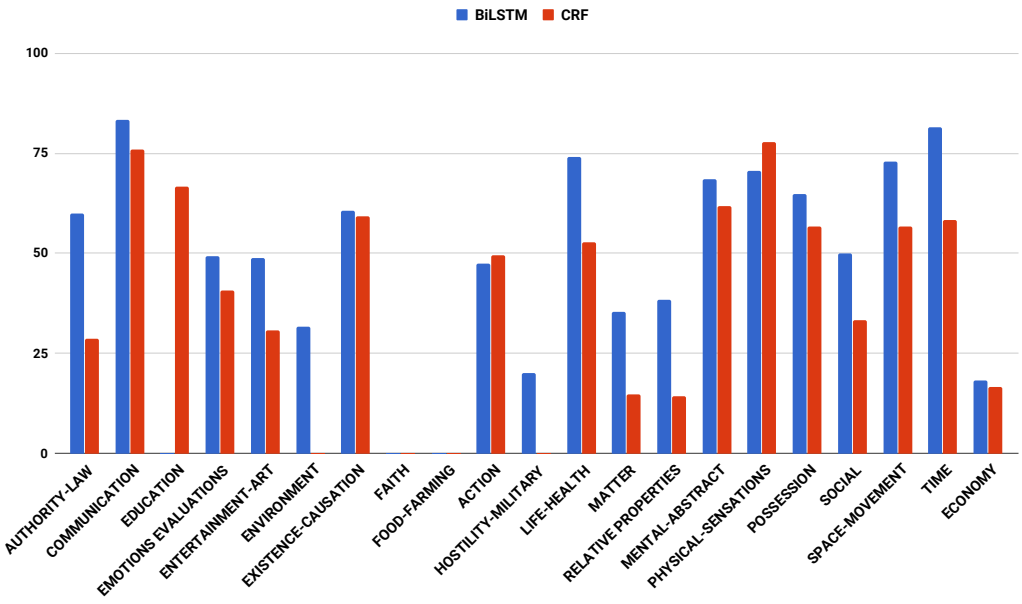

Figure 4
Comparison of F1 scores for each evaluated event class. The CLOTHES class is not in the figure because it was not present in the test set.

Compared to the results of inter-annotator agreement, we notice that three of the classes with higher F1 had also a perfect agreement between human annotators: This means that COMMUNICATION, PHYSICAL SENSATIONS, and LIFE-HEALTH are the less ambiguous classes to be identified. On the contrary, other classes with perfect IAA have very low scores or even an F1 equal to zero: this is the case for FOOD-FARMING, EDUCATION, and ECONOMY. The BiLSTM model, for example, correctly annotated only the verb "to pay" as belonging to the ECONOMY class but assigned the class RELATIVE PROPERTIES to copular constructions including monetary expressions, such as in "the loss **was $600**," interpreting these expressions as quantities. The same construction was not recognized as an event by the CRF classifier.

As for linguistic realizations, they have an impact on mention extension given that, for example, light verb constructions and phrasal verbs require a multi-token annotation. In the mention detection task both approaches perform well, with an F1 above 80%. On the contrary, differences can be seen by analyzing the combination of mention detection and class assignment in the detection+classification task. As shown in Table 12, the percentage of events for which the BiLSTM model is able to correctly annotate both the extension and the class is higher for all the linguistic realizations compared with the CRF model. It is worth noticing that fixed expressions (e.g., "had enough" in "people in the grand stand evidentally had enough of the race") and light verbs (e.g., "had misgivings" in "we had some misgivings") are the most challenging linguistic realizations.

To conclude, BiLSTM models can perform our tasks with good performance. This is particularly evident considering the task that combines both mention detection and classification, for which the CRF classifier yields much worse results.

Table 12
Accuracy in the combined detection and classification of events for different linguistic realizations.

	ACCURACY	
	BiLSTM	CRF
VERBS	70%	53%
PHRASAL VERBS	56%	45%
NOUNS	54%	34%
COPULAR CONSTRUCTIONS	46%	32%
FIXED EXPRESSIONS	36%	28%
LIGHT VERBS	25%	0%

7. Conclusions

In this work, we provided a theoretical and practical investigation on the topic of event detection and classification in historical texts. In particular, we presented new annotation guidelines for event mention detection and classification, and then a new manually annotated corpus of historical texts, the largest publicly available to address the task in this domain. Finally, we dealt with event mention detection and classification using a deep learning architecture and comparing the results with the ones achieved with a CRF classifier. An additional contribution is the thorough analysis of the impact of different word embeddings on the task, comparing both the effect of the source corpus (GoogleNews, COHA corpus) and the type of information encoded in the vectors (bag-of-words, bag-of-characters, dependency trees): Results show that dependency-based embeddings better capture the type of information needed to classify events, and a large, generic corpus like GoogleNews is still preferable over a smaller, domain-specific one to create the vectors.

The deep neural model we developed for the task including both the detection and the classification of event mention is, to all effects, an end-to-end system that can be applied to raw texts with satisfactory results, especially for some semantic classes of events such as those related to communication, motion, mental actions, and process. This system can represent the basis for the creation of a complete framework in which to integrate other NLP tools already available to the research community. For example, modules for temporal and causal relations extraction, event factuality detection, and semantic role labeling can be added on top of our system.

Acknowledgments

The authors would like to thank Giovanni Moretti for technical assistance, Anna Feltracco for her help with inter-annotator agreement, and the anonymous reviewers for their helpful comments.

References

Aguilar, Jacqueline, Charley Beller, Paul McNamee, Benjamin Van Durme, Stephanie Strassel, Zhiyi Song, and Joe Ellis. 2014. A comparison of the events and relations across ACE, ERE, TAC-KBP, and FrameNet annotation standards. In *Proceedings of the 2nd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 45–53, Baltimore, MD.

Alexander, Marc and Andrew Struan. 2013. 'In countries so unciviliz'd as those?': The language of incivility and the British experience of the world, In *The British Abroad Since the Eighteenth Century, Volume 2*, Springer, pages 232–249.

Artstein, Ron and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Atefeh, Farzindar and Wael Khreich. 2015. A survey of techniques for event detection in Twitter. *Computational Intelligence*, 31(1):132–164.
Bach, Emmon. 2008. The algebra of events. *Formal Semantics: The Essential Readings*, 9(1):324–333.
Bartalesi Lenzi, Valentina, Giovanni Moretti, and Rachele Sprugnoli. 2012. CAT: The CELCT annotation tool. In *Proceedings of LREC 2012*, pages 333–338, Istanbul.
Bethard, Steven, Leon Derczynski, Guergana Savova, James Pustejovsky, and Marc Verhagen. 2015. SemEval-2015 Task 6: Clinical TempEval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 806–814, Denver, CO.
Bethard, Steven, Guergana Savova, Wei-Te Chen, Leon Derczynski, James Pustejovsky, and Marc Verhagen. 2016. SemEval-2016 Task 12: Clinical TempEval. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1052–1062, San Diego, CA.
Bethard, Steven, Guergana Savova, Martha Palmer, and James Pustejovsky. 2017. SemEval-2017 Task 12: Clinical TempEval. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 565–572, Vancouver.
Bingham, Adrian. 2010. The digitization of newspaper archives: Opportunities and challenges for historians. *Twentieth Century British History*, 21(2):225–231.
Björne, Jari and Tapio Salakoski. 2011. Generalizing biomedical event extraction. In *Proceedings of the BioNLP Shared Task 2011 Workshop*, pages 183–191, Portland, OR.
Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.
Burke, Peter. 1997. *Varieties of Cultural History*. Cornell University Press.
Carletta, Jean. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.
Ciaramita, Massimiliano and Mark Johnson. 2003. Supersense tagging of unknown nouns in WordNet. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 168–175, Sapporo.
Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.
Collobert, Ronan, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.
Cybulska, Agata and Piek Vossen. 2010. 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 *Proceedings of LREC*, pages 3355–3362, Valletta.
Cybulska, Agata and Piek Vossen. 2011. Historical event extraction from text. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 39–43, Portland, OR.
Czeitschner, Ulrike and Barbara Krautgartner. 2017. Discursive constructions of culture: Semantic modelling for historical travel guides. *Sociology and Anthropology*, 5(4):323–331.
Davies, Mark. 2012. Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English. *Corpora*, 7(2):121–157.
De Boer, Victor, Johan Oomen, Oana Inel, Lora Aroyo, Elco Van Staveren, Werner Helmich, and Dennis De Beurs. 2015. DIVE into the event-based browsing of linked historical media. *Web Semantics: Science, Services and Agents on the World Wide Web*, 35:152–158.
Derczynski, Leon, Jannik Strötgen, Diana Maynard, Mark A. Greenwood, and Manuel Jung. 2016. Gate-time: Extraction of temporal expressions and event. In *10th Language Resources and Evaluation Conference*, pages 3702–3708, Portorož.
Dice, Lee R. 1945. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302.
den Dikken, Marcel and Teresa O'Neill. 2016. Copular constructions in syntax. In *Oxford Research Encyclopedia of Linguistics*, Interactive Factory.
Doddington, George R, Alexis Mitchell, Mark A. Przybocki, Lance A. Ramshaw, Stephanie Strassel, and Ralph M. Weischedel. 2004. The automatic content extraction (ACE) program—tasks, data, and evaluation. In *LREC*, pages 837–840, Lisbon.
Dölling, Johannes. 2014. Aspectual coercion and eventuality structure. *Events, Arguments, and Aspects. Topics in the Semantics of Verbs*, 189–226.
Dozat, Timothy. 2016. Incorporating Nesterov momentum into Adam. In

Proceedings of ICLR 2016, pages 1–4, San Juan, PR.

Duchi, John, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159.

Filatova, Elena and Eduard Hovy. 2001. Assigning time-stamps to event-clauses. In *Proceedings of the ACL-EACL 2001 Workshop for Temporal and Spatial Information Processing*, 88–95, Toulouse.

Fokkens, Antske, Serge ter Braake, Niels Ockeloen, Piek Vossen, Susan Legêne, Guus Schreiber, and Victor de Boer. 2018. Biographynet: Extracting relations between people and events. *arXiv preprint arXiv:1801.07073*.

Fokkens, Antske, Serge Ter Braake, Niels Ockeloen, Piek Vossen, Susan Legêne, and Guus Schreiber. 2014. BiographyNet: Methodological issues when NLP supports historical research. In *Proceedings of LREC*, pages 3728–3735, Reykjavik.

Guerrero Nieto, Marta, Roser Saurì, and Miguel Angel Bernabé Poveda. 2011. Modes timebank: A modern Spanish timebank corpus. *Procesamiento del Lenguaje Natural*, 47:259–267.

Hamilton, William L., Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin.

Ide, Nancy and Catherine Macleod. 2001. The American National Corpus: A standardized resource of American English. In *Proceedings of Corpus Linguistics*, 3, pages 1–7, Lancaster.

Ide, Nancy and David Woolner. 2004. Exploiting semantic web technologies for intelligent access to historical documents. In *Computer*, pages 2177–2180, Reykjavik.

Ide, Nancy and David Woolner. 2007. Historical ontologies. *Words and Intelligence II*, pages 137–152.

Iso, SemAf/Time Working Group. 2008. ISO DIS 24617-1: 2008 Language resource management. Semantic annotation framework - Part 1: Time and events. Technical report, ISO Central Secretariat, Geneva.

Jung, Hyuckchul and Amanda Stent. 2013. ATT1: Temporal annotation using big windows and rich syntactic and semantic features. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, 2, pages 20–24, Atlanta, GA.

Katz, Graham and Fabrizio Arosio. 2001. The annotation of temporal information in natural language sentences. In *Proceedings of the Workshop on Temporal and Spatial Information Processing*, 13 page 15, Toulouse.

Kay, Christian, Jane Roberts, Michael Samuels, and Irené Wotherspoon. 2009. *Historical Thesaurus of the Oxford English Dictionary*. Oxford University Press.

Kemmer, Suzanne and Arie Verhagen. 1994. The grammar of causatives and the conceptual structure of events. *Cognitive Linguistics*, 5(2):115–156.

Keskar, Nitish Shirish and Richard Socher. 2017. Improving generalization performance by switching from Adam to SGD. *CoRR*, abs/1712.07628.

Kim, Jin Dong, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. 2009. Overview of BioNLP'09 shared task on event extraction. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, pages 1–9, Boulder, CO.

Kim, Jin Dong, Tomoko Ohta, Tomoko Tateisi, and Junichi Tsujii. 2006. GENIA corpus manual, Tsujiilab, University of Tokyo.

Kim, Jin Dong, Sampo Pyysalo, Tomoko Ohta, Robert Bossy, Ngan Nguyen, and Jun'ichi Tsujii. 2011. Overview of BioNLP shared task 2011. In *Proceedings of the BioNLP Shared Task 2011 Workshop*, pages 1–6, Portland, OR.

Kim, Jin Dong, Yue Wang, Nicola Colic, Seung Han Beak, Yong Hwan Kim, and Min Song. 2016. Refactoring the Genia event extraction shared task toward a general framework for IE-driven KB development. In *Proceedings of the 4th BioNLP Shared Task Workshop*, pages 23–31, Berlin.

Kingma, Diederik and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kolomiyets, Oleksandr and Marie-Francine Moens. 2013. KUL: A data-driven approach to temporal parsing of documents. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics (* SEM)*, 2, pages 83–87, Atlanta, GA.

Komninos, Alexandros and Suresh Manandhar. 2016. Dependency based

embeddings for sentence classification tasks. In *HLT-NAACL*, pages 1490–1500, San Diego, CA.
Lample, Guillaume, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, CA.
Landis, J. Richard and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174.
Le Boeuf, Patrick, Martin Doerr, Christian Emil Ore, and Stephen Stead. 2017. Definition of the CIDOC conceptual reference model, Version 6.2.2. Technical report, ICOM/CIDOC Documentation Standards Group. CIDOC CRM Special Interest Group.
Lee, Hee Jin, Hua Xu, Jingqi Wang, Yaoyun Zhang, Sungrim Moon, Jun Xu, and Yonghui Wu. 2016. UTHealth at SemEval-2016 Task 12: an End-to-End System for Temporal Information Extraction from Clinical Notes. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1292–1297, San Diego, CA.
Lever, Jake and Steven J. M. Jones. 2016. VERSE: Event and relation extraction in the BioNLP 2016 Shared Task. In *Proceedings of the 4th BioNLP Shared Task Workshop*, pages 42–49, Berlin.
Levy, Omer and Yoav Goldberg. 2014. Dependency-based word embeddings. In *ACL (2)*, pages 302–308.
Li, Chen, Zhiqiang Rao, and Xiangrong Zhang. 2016. LitWay, discriminative extraction for different bio-events. In *Proceedings of the 4th BioNLP Shared Task Workshop*, pages 32–41.
Linguistic Data Consortium. 2005. ACE (Automatic Content Extraction) English annotation guidelines for events, Technical Report version 5.4.3 2005.07.01, LDC.
Llorens, Hector, Estela Saquete, and Borja Navarro. 2010. TIPSem (English and Spanish): Evaluating CRFs and semantic roles in TempEval-2. In *ACL 2010 - SemEval 2010 - Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 284–291, Association for Computational Linguistics, Uppsala, Sweden.
Ma, Xuezhe and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNS-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin.
Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *ACL (System Demonstrations)*, pages 55–60.
Mauri, Michele, Tommaso Elli, Giorgio Caviglia, Giorgio Uboldi, and Matteo Azzi. 2017. Rawgraphs: A visualisation platform to create open outputs. In *Proceedings of the 12th Biannual Conference on Italian SIGCHI Chapter*, pages 28, Cagliari.
McNally, Louise. 1998. Existential sentences with existential quantification. *Linguistics and Philosophy*, 21(4):353–392.
Mehryary, Farrokh, Jari Björne, Sampo Pyysalo, Tapio Salakoski, and Filip Ginter. 2016. Deep learning with minimal training data: TurkuNLP entry in the BioNLP shared task 2016. In *Proceedings of the 4th BioNLP Shared Task Workshop*, pages 73–81, Berlin.
Mikkelsen, Line. 2005. *Copular clauses: Specification, predication and equation*, 85. John Benjamins Publishing.
Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
Minard, Anne Lyse, Manuela Speranza, Eneko Agirre, Itziar Aldabe, Marieke van Erp, Bernardo Magnini, German Rigau, Rubén Urizar, and Fondazione Bruno Kessler. 2015. SemEval-2015 Task 4: TimeLine: Cross-document event ordering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 778–786, Denver, CO.
Mitamura, Teruko, Zhengzhong Liu, and Eduard Hovy. 2016. Overview of TAC-KBP 2016 Event Nugget Track. In *Proceedings of the Ninth Text Analysis Conference*. Gaithersburg, MD, https://tac.nist.gov/publications/2015/papers.html.
Mitamura, Teruko, Zhengzhong Liu, and Eduard Hovy. 2017. Events detection, coreference and sequencing: What's next? Overview of the TAC KBP 2017 event track. In *Proceedings of the Text Analysis Conference*, Gaithersburg, MD. https://tac.nist.gov/publications/2017/additional.paper/TAC2017.KBP_Event_Nugget_overview.proceedings.pdf
Mitamura, Teruko, Yukari Yamakawa, Susan Holm, Zhiyi Song, Ann Bies, Seth Kulick, and Stephanie Strassel. 2015. Event Nugget Annotation: Processes and issues.

In *Proceedings of the 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 66–76, Denver, CO.

Nédellec, Clair, Philippe Bessieres, Robert Bossy, Alain Kotoujansky, and Alain-Pierre Manine. 2006. Annotation guidelines for machine learning-based named entity recognition in microbiology. In *Proceedings of the ACL Workshop on Data and Text for Mining Integrative Biology*, pages 40–54, Berlin.

Nédellec, Claire, Robert Bossy, Jin-Dong Kim, Jung-jae Kim, Tomoko Ohta, Sampo Pyysalo, and Pierre Zweigenbaum. 2013. Overview of BioNLP shared task 2013. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 1–7, Sofia.

Oakeshott, Michael. 2015. *Experience and its Modes*. Cambridge University Press.

Pennington, Jeffrey, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha.

Piao, Scott, Fraser Dallachy, Alistair Baron, Jane Demmen, Steve Wattam, Philip Durkin, James McCracken, Paul Rayson, and Marc Alexander. 2017. A time-sensitive historical thesaurus-based semantic tagger for deep semantic annotation. *Computer Speech & Language*, 46:113–135.

Pustejovsky, James. 1991. The generative lexicon. *Computational Linguistics*, 17(4):409–441.

Pustejovsky, James, José Castano, Robert Ingria, Roser Saurí, Robert Gaizauskas, Andrea Setzer, and Graham Katz. 2003. TimeML: Robust specification of event and temporal expressions in text. In *Proceedings of IWCS-5*, pages 1–12, Tilburg.

Raimond, Yves and Samer Abdallah. 2007. The event ontology, Citeseer.

Reimers, Nils and Iryna Gurevych. 2015. Event nugget detection, classification and coreference resolution using deep neural networks and gradient boosted decision trees. In *Proceedings of the Eight Text Analysis Conference (TAC 2015)*, Gaithersburg, MD.

Reimers, Nils and Iryna Gurevych. 2017a. Optimal hyperparameters for deep LSTM-networks for sequence labeling tasks. *arXiv preprint arXiv:1707.06799*.

Reimers, Nils and Iryna Gurevych. 2017b. Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 338–348, Copenhagen.

Robbins, Herbert and Sutton Monro. 1951. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3), 400–407.

Roberts, Jane. 2000. The 'historical thesaurus' as a tool for medievalists. *Studies in Modern English*, 2000(16):1–23.

Sablayrolles, Pierre. 1995. The semantics of motion. In *Proceedings of the Seventh Conference on European Chapter of the Association for Computational Linguistics*, pages 281–283, Cambridge, MA.

Sang, Erik F. and Jorn Veenstra. 1999. Representing text chunks. In *Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics*, pages 173–179, Singapore.

Schilder, Frank and Christopher Habel. 2003. In Temporal information extraction for temporal question answering, *New Directions in Question Answering*, pages 35–44.

Segers, Roxane, Marieke Van Erp, Lourens Van Der Meij, Lora Aroyo, Guus Schreiber, Bob Wielinga, Jacco van Ossenbruggen, Johan Oomen, and Geertje Jacobs. 2011. Hacking history: Automatic historical event extraction for enriching cultural heritage multimedia collections. In *Proceedings of the 6th International Conference on Knowledge Capture (K-CAP'11)*, pages 26–29, Banff.

Shaw, Ryan, Raphaël Troncy, and Lynda Hardman. 2009. LODE: Linking open descriptions of events. *ASWC*, 9:153–167.

Shaw, Ryan Benjamin. 2010. *Events and Periods as Concepts for Organizing Historical Knowledge*. University of California, Berkeley.

Sohn, Sunghwan, Sean P. Murphy, James J. Masanz, Jean-Pierre A. Kocher, and Guergana K. Savova. 2010. Classification of medication status change in clinical narratives. In *AMIA Annual Symposium Proceedings*, 2010, pages 762–766, Washington, DC.

Song, Zhiyi, Ann Bies, Stephanie Strassel, Joe Ellis, Teruko Mitamura, Hoa Dang, Yukari Yamakawa, and Sue Holm. 2016. Event nugget and event coreference annotation. In *Proceedings of 4th Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 37–45, San Diego, CA.

Song, Zhiyi, Ann Bies, Stephanie Strassel, Tom Riese, Justin Mott, Joe Ellis, Jonathan Wright, Seth Kulick, Neville Ryant, and Xiaoyi Ma. 2015. From light to rich ERE: Annotation of entities, relations, and events. In *Proceedings of the 3rd Workshop on EVENTS at the NAACL-HLT*, pages 89–98, Denver, CO.

Speranza, Manuela and Rachele Sprugnoli. 2018. Annotation of temporal information on historical texts: A small corpus for a big challenge. In Cotticelli-Kurras, Paola and Federico Giusfredi, editors, *Formal Representation and Digital Humanities*, Cambridge Scholars Publishing, Cambridge, pages 212–229.

Sprinthall, Richard C. 2003. *Basic Statistical Analysis*. Allyn & Bacon.

Sprugnoli, Rachele. 2018. Event Detection and Classification for the Digital Humanities. Ph.D. thesis, University of Trento.

Sprugnoli, Rachele. 2018b. "Two days we have passed with the ancients...": A digital resource of historical travel writings on italy. In *AIUCD2018 - Book of Abstracts*, pages 242–245.

Sprugnoli, Rachele and Sara Tonelli. 2017. One, no one, and one hundred thousand events: Defining and processing events in an inter-disciplinary perspective. *Natural Language Engineering*, 23(4):485–506.

Sprugnoli, Rachele, Sara Tonelli, Giovanni Moretti, and Stefano Menini. 2017. A little bit of bella pianura: Detecting code-mixing in historical English travel writing. In *Fourth Italian Conference on Computational Linguistics (CLiC-it 2017)*, pages 304–309, Rome.

Styler, William F. IV, Guergana Savova, Martha Palmer, James Pustejovsky, Tim O'Gorman, and Piet C. de Groen. 2014. Thyme annotation guidelines. Technical report, University of Colorado at Boulder.

Sun, Weiyi, Anna Rumshisky, and Ozlem Uzuner. 2013. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. *Journal of the American Medical Informatics Association*, 20(5):806–813.

Sundheim, Beth M. 1991. Overview of the third message understanding evaluation and conference. In *Proceedings of the Third Message Understanding Conference (MUC-3)*, pages 3–16, San Diego, CA.

Talmy, Leonard. 1996. Fictive motion in language and 'ception'. *Language and Space*, 21:1–276.

Teufel, Simone, Advaith Siddharthan, and Colin Batchelor. 2009. Towards discipline-independent argumentative zoning: Evidence from chemistry and computational linguistics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, pages 1493–1502, Singapore.

The European Union. 2012. Definition of the Europeana Data Model elements. Europeana.

Tieleman, Tijmen and Geoffrey Hinton. 2012. Lecture 6.5-rmsprop, coursera: Neural networks for machine learning. University of Toronto, Technical Report.

Tourille, Julien, Olivier Ferret, Xavier Tannier, and Aurélie Névéol. 2017. LIMSI-COT at SemEval-2017 Task 12: Neural architecture for temporal information extraction from clinical narratives. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 597–602,Vancouver.

UzZaman, Naushad, Hector Llorens, Leon Derczynski, Marc Verhagen, James Allen, and James Pustejovsky. 2013. Semeval-2013 task 1: {T}empeval-3: Evaluating time expressions, events, and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (* SEM)*, 2, pages 1–9, Atlanta, GA.

Van Hage, Willem Robert, Véronique Malaisé, Roxane Segers, Laura Hollink, and Guus Schreiber. 2011. Design and use of the simple event model (SEM). *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(2):128–136.

Velupillai, Sumithra, Danielle L. Mowery, Samir Abdelrahman, Lee Christensen, and Wendy Chapman. 2015. BluLab: Temporal information extraction for the 2015 clinical TempEval challenge. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 815–819, Denver, CO.

Vendler, Zeno. 1957. In Verbs and times, *The Philosophical Review*, 66:143.

Verhagen, Marc, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. 2007. SemEval-2007 Task 15 : TempEval temporal relation identification. In *Computational Linguistics*, pages 75–80, Prague.

Verhagen, Marc, Roser Saurí, Tommaso Caselli, and James Pustejovsky. 2010. SemEval-2010 task 13: TempEval-2. In *ACL 2010 - SemEval 2010 - 5th International*

Workshop on Semantic Evaluation, Proceedings, pages 57–62, Uppsala.

Zavarella, Vanni and Hristo Tanev. 2013. FSS-TimEx for TempEval-3: Extracting temporal information from text. In *Proceedings of SemEval 2013*, pages 58–63, Atlanta, GA.

Zeiler, Matthew D. 2012. ADADELTA: An adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

Zeldes, Amir. 2017. The gum corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.