



Event-based Access to Historical Italian War Memoirs

MARCO ROVERA, Department of Computer Science, University of Torino, Italy and Institute for Applied AI (IAAI), Stuttgart Media University, Germany

FEDERICO NANNI and SIMONE PAOLO PONZETTO, Data and Web Science Group (DWS), University of Mannheim, Germany

The progressive digitization of historical archives provides new, often domain-specific, textual resources that report on facts and events that have happened in the past; among these, memoirs are a very common type of primary source. In this article, we present an approach for extracting information from Italian historical war memoirs and turning it into structured knowledge. This is based on the semantic notions of events, participants, and roles. We evaluate quantitatively each of the key steps of our approach and provide a graph-based representation of the extracted knowledge, which allows to move between a Close and a Distant Reading of the collection.

CCS Concepts: • **Computing methodologies** → **Information extraction**; Lexical semantics; • **Applied computing** → Document management and text processing;

Additional Key Words and Phrases: Digital history, event extraction, entity linking, distant reading, second world war

ACM Reference format:

Marco Rovera, Federico Nanni, and Simone Paolo Ponzetto. 2021. Event-based Access to Historical Italian War Memoirs. *J. Comput. Cult. Herit.* 14, 1, Article 2 (January 2021), 23 pages.
<https://doi.org/10.1145/3406210>

1 INTRODUCTION

The growing interest of cultural institutions in the digitization of archival documents and resources, along with the availability of new, born-digital textual materials about historical topics, raises the question of how to provide the users with an account of the knowledge contained in such collections. From a computational point of view, one of the challenges lies in the capability of developing models and techniques to automatically extract information from historical (most of the time OCR-digitized) documents and turn such information into knowledge that can be easily accessed, queried, and visualized.

Authors' addresses: M. Rovera, Department of Computer Science, University of Torino, Corso Svizzera, 185, Torino, Italy, 10149, Institute for Applied AI (IAAI), Stuttgart Media University, Nobelstrasse, 10, Stuttgart, Germany, 70569; email: rovera@hdm-stuttgart.de; F. Nanni and S. P. Ponzetto, Data and Web Science Group (DWS), University of Mannheim, B6, 26, Mannheim, Germany; emails: {federico, simone}@informatik.uni-mannheim.de.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2021 Copyright held by the owner/author(s). Publication rights licensed to ACM.

1556-4673/2021/01-ART2 \$15.00

<https://doi.org/10.1145/3406210>

The possibility to deal with thousands of books or documents at once [11] was the main foreground for the idea of “Distant Reading” [33, 34]. The Distant Reading approach suggested a new paradigm based on the application of computational tools for analyzing large literary collections, fostering a new approach for scholars in the humanities methodologically closer to the social sciences.

During the last two decades, with the growing availability of applications following the general framework of Distant Reading in historical research [18, 34], several examples of advanced access to digitized collections through the use of text-mining technologies have been presented (see for instance the works by Blevins [6], Kaufman [25], Wilkens [61]). Though results have been fascinating, the Distant Reading paradigm has been also strongly criticized, especially for its bias towards quantification and scientification of humanities’ research problems [44]. Moreover, while nowadays established Distant Reading approaches such as Topic Modeling and Named Entity Recognition support the users in going beyond traditional keyword searches, often these text-mining techniques produce only coarse-grained macro-overviews of the information contained in the collection under study, for instance by providing a list of the most frequently mentioned entities or the most recurrent topics (as already discussed by Jänicke et al. [23], Nanni et al. [36]). Additionally, these methods generally lack the possibility of rendering back semantic information in a more fine-grained way (e.g., by retrieving highly relevant sentences). The possibility of moving between (macro-) quantitative analysis and (micro-) qualitative studies is now a relevant topic both in Digital Humanities and Computational Linguistics, especially because it could allow to get a better understanding of the model, the corpus, and the underlying assumptions.

With this in mind, the goal of our work is to propose an event-based Distant Reading approach that additionally allows to obtain a closer, semantic, and content-aware reading of the sources behind the creation of each specific pattern in the output. This, to (a) overcome the previously mentioned criticism that Distant Reading approaches keep the reader too far away from the source [54], (b) offer a better understanding of how our method works, and (c) recognize the value of both quantitative and qualitative types of approaches to historical sources [23]. In Section 7, we discuss the advantages of such representation in a series of scenarios that allow both Close and Distant Reading of the corpus. The long-term aim of this effort is to support scholars and other stakeholders in Digital and Public History to dive into large textual collections and to fulfill their information needs by exploring many different textual sources in a consistent way. To do so, we develop a pipeline for extracting events and event participants, along with their roles, from text; additionally, we resolve Named Entity co-reference to be able to match each unique entity to the set of events it participates in. Finally, we test this pipeline in a real-world scenario on a corpus of Italian war memoirs concerning the Second World War, digitized for this purpose. We present an extensive evaluation and error analysis of each step of our pipeline.

Outline. The rest of the article is structured as follows: In Section 2, we discuss some of the most relevant previous works related to our study from the fields of Text Mining and Event Extraction. Section 3 describes the textual corpus and the knowledge resources used in our system. Section 4 presents the way we model events in our work, which led to the creation of a lexical resource for extraction of events and participants. In Section 5 the extraction pipeline is sketched in all its components, and in Section 6 evaluation results are discussed for each of its key steps. In Section 7, we discuss some practical use case scenarios for event-based access to the extracted knowledge base; this allows moving from the textual collection to a network representing events, participants, and roles, back again to the text. Section 8 concludes the article and briefly discusses the reusability of our solution and resources by other researchers, along with the future directions we envisage for our work. Finally, Appendix A contains the list of digitized historical texts we employ in this work. The gazetteers (4.2), the event-predicate dictionary (4.2), the final event data (5.3), and a Gephi event-entity graph (7), together with code to process them are available on github.¹

¹github.com/marcrovera/ita-resistance.

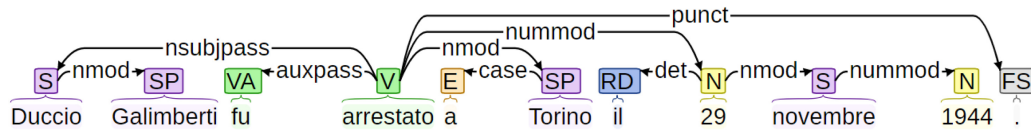


Fig. 1. Example of the correspondence between syntactic arguments of the verb and participants of the event denoted by the verb (translation: *Duccio Galimberti was arrested in Torino on the 29th November 1944*). This image has been created using the Simpatico project demo, available at <http://simpatico.fbk.eu/demo/>.

2 RELATED WORKS

In this section, we first offer an overview of related works on advanced access to digital library collections through the use of text-mining methods. Next, we cover previous research on Information Extraction and Semantic Role Labeling, relevant for our work.

Advanced Access to Textual Collections. During the past 15 years Latent Dirichlet Allocation (LDA) topic modeling [5, 40] has arguably been the most popular text-mining technique for corpus exploration in Digital Humanities (DH) and Digital Libraries (DL). This approach has been adopted to offer advanced access to scientific collections [28], proceedings of political debates [19], and historical corpora [39]. To face the limitations of LDA topics [10] and to extract information easier to interpret for final users, these communities have seen, in recent years, a growth in the combination of Named Entity Recognition (and whenever possible linking) with network analysis techniques. Kaufman [25] relies on these approaches for examining the Digital National Security Archive (DNSA) Kissinger Collections, while Menini et al. [30] use them for tracing the movements of popular historical figures mentioned in Wikipedia and Ardanuy and Sporleder [3] for clustering novels by genres and authors.

Following the potential of such combination of Named Entity Recognition and network analysis technologies for offering advanced access to historical collections, in this article, we intend to go a few steps deeper into the semantic information that could be extracted from textual data. To do so, we identify and disambiguate mentions of entities with a domain-specific knowledge resource, tag them with the specific semantic role (depending on the event in which they appear), and highlight their network of relations.

Event Extraction. This article focuses on the extraction of fine-grained events for providing advanced access to historical collections. The choice of events as a pivotal notion is motivated by the consideration that events are a “natural” structuring concept, at both a representation and a linguistic level, as they tie together time, space, and participants; this observation holds true especially when dealing with historical texts. Events have been the focus of many different research areas, notably Computational Ontology, Computational Linguistics, and Natural Language Processing, among others. Due to the magnitude of this research, we restrict our focus on recent interest from the Digital Library community in event-based collection building [17, 24, 37, 38, 41] and its applications. In NLP and related areas, the notion of event has been modeled and applied to different sub-tasks, such as Question Answering [63], Topic Detection and Tracking [1, 26], Narrative Chain Induction [8, 9], Entity Disambiguation [46], and Information Extraction and Retrieval [55, 56]. Applications in this area are well exemplified by a research thread focusing on different aspects of events in historical texts and newswires, for example Extraction [12, 52] and Coreference Resolution [13, 14]. This thread was also part of NewsReader [60], a multilingual, EU-funded research project focused on Event and Information Extraction from newswires. Our effort, especially the design choice of focusing on fine-grained representations of events, is further encouraged by the findings provided in their work by Sprugnoli and Tonelli [57]; in a multidisciplinary setting, the authors investigate the gap between computational techniques and the community of historians. They directly involve scholars from this area by conducting a survey, aimed at understanding how events are envisioned by the research community. One of the findings reported by the authors is that the fine-grained type of an event is considered important by historians, but existing annotation schemas (such as ACE [15], Event Nugget [32], or

Table 1. Composition of the Textual Collection

	Documents	Words	Sentences	Acquisition
Memoirs	25 books	1,469,000	95,500	Digitization
Bio	189 entries	57,400	2,500	Web scraping
Wiki	1,748 articles	1,364,000	53,800	Web scraping
Mem-test	3 books	112,000	5,100	Digitization

TimeML [45, 49]) mostly provide broad event classifications. Building upon this previous study, our work aims at providing a real-world application of entity and event detection techniques on a newly digitized historical collection of war memoirs.

Semantic Role Labeling and Frame Semantics. Semantic Role Labeling (SRL) is the task of automatically assigning a semantic role to a portion of text in a sentence. A *semantic role* is a label describing the thematic role played by a word or a group of words with respect to the main action or state described in the sentence; for example an AGENT carries out an action, a PATIENT is subjected to an action or to its consequences, a SOURCE marks the starting point of a movement. Since it requires a pre-defined set of semantic roles, along with a linguistic knowledge base in which relationships between lexical units (i.e., words) and semantic roles are formally made explicit, SRL is generally approached as a supervised task. Typical lexical knowledge bases for SRL in English are FrameNet [4], VerbNet [50], and Propbank [43]. Since such knowledge bases exist only partially for Italian, in this work, we created a lightweight resource, focusing on three high-level event categories: movements, conflict events, and social events concerning membership and organizations.

3 TEXTS AND RESOURCES

In this section, we provide an overview of the collections and the knowledge resources employed in our work.

3.1 Textual Collections

The textual collection employed in this work is divided into three subcorpora, as presented in Table 1. All textual resources are in Italian. The central resource comprises historical war memoirs of Italian partisans from World War II in North-Western Italy (Memoirs and Memoirs-test). This is accompanied by two other related resources: biographic records (Biographies) of the participants to these events and encyclopedic entries on the topic (Wiki-Articles). The Memoirs subcorpus is the one of interest for the scope of this work and the one that the system will be tested directly on. We consider the availability of such textual corpus of great importance for many reasons. In fact, we claim that one of the responsibilities of Digital History is to preserve and make accessible the past from multiple perspectives, by paying attention both to the individual and collective stories of women and men who acted in the past and to their relationship and role within the great events that “make the history.” Moreover, our texts are one example of textual resources that are typically provided by historical archives and, as non-digital-born resources, they challenge current computational techniques that are often based on and oriented to digital sources (e.g., Wikipedia, Twitter) or standard texts (e.g., newspapers). The Biographies and the Wiki-Articles subcorpora have been employed for expanding the lexical resource for Event Extraction due to the fact that, while they are written in different styles, they deal with the same topic. This guarantees a higher lexical variety and a wider coverage of the relation between argument structures and denoted event mentions.

Memoirs. This subcorpus consists of 25 books, historical memoirs of Italian partisans from World War II in North-Western Italy. The time span of the depictions ranges from the 8 September, 1943, to 25 April, 1945, a period known in the Italian historiography as “Resistenza” (Resistance). Out of 25 books, 20 have been obtained by manual digitization from the original printed editions, while the remaining 5 have been acquired through

automatic conversion from existing digital editions. The digitization has been performed first by scanning the original sources and then by automatic conversion to text using an OCR software (Adobe Acrobat Pro DC, version 2015). Despite the good performance of the employed OCR, a subsequent manual cleaning has been necessary. The definition of the corpus, digitization and manual cleaning thereof required about two man-months. This acquisition effort resulted in a textual corpus of $\approx 1.5\text{M}$ words and over 95,000 sentences.

Biographies. The second subcorpus has been obtained from the Wikisource page dedicated to “Women and Men of the Italian Resistance,”² which contains short biographic records of over 3,000 persons involved in the Italian Resistance provided by the National Association of Italian Partisans (ANPI). We kept all entries related to persons who also appear in the Memoirs corpus, for a total of 189 biographies, 57,400 words, and 2,500 sentences.

Wiki-Articles. The last subcorpus has been created by collecting 1,748 articles from the Italian Wikipedia, corresponding to the category “Resistance movements during World War II.”³ This subcorpus counts $\approx 1.3\text{M}$ words and 53,000 sentences. The category groups together all the entries connected to resistance movements against Nazi fascism during World War II, so the Italian Resistance movement (of which the Memoirs subcorpus only covers a minor part) is a subset of it. Additionally, the Wiki subcorpus is far more heterogeneous than the previous one, as the documents span over a range of thematic categories (literature, cinema, historical events).

Memoirs-test. To evaluate the system and the event-predicate dictionary (see Section 3.2) on the task of Event Extraction, a test corpus has been created employing memoirs that we have initially excluded from the Memoirs collection. It consists of 112,000 words and 5,100 sentences.

3.2 Knowledge Resources

Beside the textual corpus, the pipeline for Event Extraction requires a set of knowledge resources; each of them is used in our work for a specific task. Below, we provide an overview of the employed knowledge resources and clarify their role in the pipeline, which will be described in more detail in Section 5.

Gazetteers. The majority of the books belonging to the Memoirs sub-corpus are accompanied by a list of names of the people mentioned. The partisans are often paired with their related nickname, or *nom de guerre*. We manually merged these lists and, during the digitization of the corpus, we extended them to consider also the specific Locations and Organizations present in the text (with their related abbreviations and/or acronyms). As we have already shown in a previous work [47], employing a specific knowledge resource for this task is necessary due to the lack of domain coverage of general purpose knowledge bases derived from Wikipedia. The gazetteers are used in the pipeline for extraction and disambiguation of Named Entities and consist of 3,041 Persons, 1,725 Locations, and 245 Organizations.

Semantic Types Dictionary. Based on our background knowledge of the domain, a set of representative nominal words is assigned to each of the 25 Semantic Types (see Section 4 for more details). The number of words for each type varies, ranging from 68 for the HUMAN_COLLECTIVE type to 1 for the WATERFLOW type. This variation also accounts for the lexical variety and richness encoded by each category. For example, the HUMAN semantic type represents not only common words designating persons, such as *signora*, *uomo*, or *ragazza*,⁴ but also words referring to social roles, such as *capitano*, *madre*, *attivista*, *carceriere*,⁵ which denote human beings through the role they have in a social context. This small lexical resource is composed of 424 words and is employed in the semantic tagging step, discussed in Section 5.

²https://it.wikisource.org/wiki/Donne_e_Uomini_della_Resistenza.

³https://it.wikipedia.org/wiki/Categoria:Movimenti_di_resistenza_della_seconda_guerra_mondiale.

⁴“lady,” “man,” “girl.”

⁵“captain,” “mother,” “activist,” “jailer.”

Table 2. Example of
Syntactic-semantic Argument
Patterns Associated to the Lexical
Unit *Bombardare* (to Bomb)

bombardare (to bomb)

nsubj :: ORG

nsubj :: HUMAN_COLLECTIVE

nsubjpass :: LOC

dobj :: INFRASTRUCTURE

nmod :: [da [VEHICLE]]

nmod :: [con [WEAPON]]

A dictionary of event-predicate lexical patterns. The most relevant knowledge resource for the aim of this work is the dictionary of event-predicate lexical patterns, exemplified in Tables 2 and 3. The main purpose of the dictionary is to enable, in a single step, the recognition and classification of events, the extraction of event participants, and the labeling of each detected participant with the semantic role it plays in the event. This resource is partially inspired by the Corpus Pattern Analysis (CPA) methodology [20, 21], which aims at assigning a sense to a verbal or nominal root using a syntactic-semantic pattern of its arguments, derived from corpus evidence. On the other side, the dictionary is also inspired by FrameNet, in that it provides a way to link a set of linguistic terms to a concept (events, in our case) and to retrieve the structure of such concept in texts. As opposed to CPA, (a) our event-predicate dictionary is not meant to map lexical senses but event mentions denoted by linguistic terms, (b) each argument pattern is described in isolation and later associated to an event class in which it can appear, and (c) a syntactic-semantic argument pattern for a given term may belong to (or denote) different event classes. For example, the pattern *abbandonare*⁶ :: subject :: PER could belong to at least two event classes, namely, DEPARTING and QUIT_GROUP; depending on the denoted event, the same argument pattern induces two different semantic roles (MOVER and MEMBER, respectively). Therefore, for each annotated argument pattern, the resource also provides the semantic role typically associated with it for the given event class. The resource has been created through manual analysis of the syntactic-semantic argument structures of a given set of lexical units in the whole corpus (Memoirs, Biographies, Wiki-Articles). The event-predicate dictionary counts 246 verbal or nominal roots, mapped to 88 event classes: 124 verbs (such as *aderire*, *imprigionare*, *partire*⁷), 77 nouns (e.g., *arresto*, *liberazione*, *arrivo*⁸), 45 multiword verbal expressions (*aprire il fuoco*, *fare prigioniero*⁹). The 88 event classes cover three high-level event categories: conflict events, movements of persons and artifacts, and events concerning the membership of individuals to organizations.

4 MODELING EVENTS

Events, along with their participants, are the conceptual notion that we model to recognize their mentions in texts. Given the immense relevance that events have in the domain of historical research [57], in this work, we decided to approach three different types, namely, spatial events (movements), conflict events, and membership in organizations. The setting is inspired by FrameNet (FN) [4, 48], with the notable difference that we are only interested in modeling event information, while many FN Frames also model other type of information.¹⁰

⁶“to leave.”

⁷“to join” (an organization), “to imprison,” “to depart.”

⁸“arrest,” “liberation/release,” “arrival.”

⁹“to open fire,” “to take prisoner.”

¹⁰From an ontological point of view, Frames describe “types of situations” [42].

Table 3. Excerpts from the Mapping of Two Lexical Units to the Correspondent Event Classes

Lexical Unit	Event Class	Syn-Sem pattern	Semantic Role
<i>abbandonare</i> (to leave)	DEPARTING	nsubj :: PER	MOVER
		dobj :: LOC	SOURCE
		nsubj :: ORGANIZAT.	MOVER
		nsubj :: GROUP	MOVER
	QUIT_GROUP	nsubj :: PER	MEMBER
			dobj :: ORGANIZAT.
<i>abbattere</i> (to tear down / to kill)	DESTROYING	nsubjpass :: INFRASTR.	PATIENT
		dobj :: INFRASTR.	PATIENT
	DOWNING	nsubjpass :: VEHICLE	PATIENT
		dobj :: VEHICLE	PATIENT
		nsubj :: HUMAN	AGENT
		nsubj :: ORG	AGENT
	KILLING	nsubj :: PER	KILLER
		nsubjpass :: PER	VICTIM
		dobj :: PER	VICTIM
		nmod :: [con [WEAP.]]	INSTRUM.
nsubj :: ORG		KILLER	

In each event class, all syntactic-semantic argument patterns associated to the verb are associated to a semantic role.

With this in mind, in our work, we denote a textual event mention by the following combination of features:

- (1) a lexical unit (LU);
- (2) a set of syntactic dependencies associated to the LU (subject, object, etc.);
- (3) a set of possible semantic types as fillers of each dependency (e.g., HUMAN_COLLECTIVE, PLACE, VEHICLE, etc.);
- (4) a set of possible semantic roles that can be assigned to each combination of LU—dependency—semantic type (for example MOVER, SOURCE, VICTIM, etc.);
- (5) an event class that identifies the specific event type.

The basic assumption is that the syntactic argument structure of an LU, at least for verbal and multi-word verbal expressions, is the main source of information about the participants of an event denoted by that LU. It follows that there is a correspondence between the fillers of certain syntactic dependencies of an LU and the participants to the event. If we consider the example in Figure 1: the main verb *arrestare* (to arrest) is the head of three syntactic dependencies, each of which is the head of a phrase representing a participant in the denoted event. Thus, we obtain:

Duccio Galimberti	nsubj	PER	ARRESTED
Torino	nmod	[a [LOC]]	PLACE
29 novembre 1944	nummod	DATE	TIME

1. Lexical Units. The term lexical unit is used in FrameNet to describe a word whose occurrence in text triggers a given frame. Although from a linguistic point of view events can be referenced in text by words belonging to different parts of speech [49], in this work, we reduce the focus only to verbs, nouns, and multi-word verbal expressions. This choice is motivated by the evidence that these categories of words have explicit syntactic relations (dependencies) that provide a first basic structure for the referenced event and participants. As opposed to this, other lexical categories (notably adjectives and adverbs), which often express entity properties, do not offer this syntactic richness in Italian. Like in FrameNet, each LU can map one or more event types.

2. Syntactic Dependencies. For events denoted by verbs and by multi-word verbal expressions, we consider the active/passive subject (nsubj, nsubjpass types in Universal Dependencies¹¹), the direct object (dobj), all the nominal modifiers (nmod), which map indirect complements like temporal, spatial, and others and the numeric modifiers (nummod), usually referring to temporal or quantified complements. For nouns, only nmod dependencies are considered. Although this set of dependency types does not account for all the possible pieces of information related to the event mentioned in a sentence (especially where the sentence is constituted by more than one clause), it represents a stable syntactic structure that conveys most of the relevant information.

3. Semantic Types. Semantic types are used to label the head of lexical arguments and provide information about the entity that appears as filler of a given syntactic dependency.¹² While some types are more general (HUMAN, HUMAN_COLLECTIVE, POPULATED_PLACE), others are specific to the domain under study (POST, WEAPON). In addition, named entity types PER, LOC, and ORG are included as semantic types, as well as the four TIMEX3 temporal tags DATE, TIME, DURATION, and SET.¹³ By means of these three components (LU, syntactic dependency, semantic type) it is possible to build syntactic-semantic patterns associated to an LU (see Tables 2 and 3 for examples).

4–5. Semantic Roles and Event Classes. We employ as event classes a set of FrameNet frames and as semantic roles the corresponding Frame Element labels. We extend both of them to fully describe the collection under study and account for relevant event types for the domain at hand. We use overall 88 event classes (as described in Section 3.2) of which 52 correspond to existing FrameNet frames, while 36 are specific for modeling event types present in our collection. The latter concern in particular (a) events that are not modeled by any FrameNet frame, for instance specific war-related event classes such as DEPORTATION, LIBERATION, or AIRDROP as well as fine-grained movements, such as GET_OFF_VEHICLE or ENTER_BUILDING; (b) events that are modeled by some FrameNet frame but in a very general way, which would not account for its relevance in our corpus, for instance DISARM, modeled in FN by the EMPTYING frame, the BOMBING event, modeled in FN as an ATTACK frame, or still the RETREAT event, modeled generally as QUITTING_A_PLACE. From the perspective of the overall goal of this work, i.e., providing fine-grained access to a historical collection of war memoirs, this is the most important step, as it allows us to assess which type of event is mentioned in a given context, to associate a set of entities (Named Entities or other types of entities) to that event, and to assess which role is played, with respect to the event, by each event participant (example in Table 3).

¹¹<http://universalddependencies.org/it/dep/>.

¹²The set of Semantic Types considered in this work is depicted in Figure 3.

¹³PER, LOC, and ORG are standard named entity labels derived by using the TINT pipeline [2]. TIMEX3 labels are established temporal tags used for instance by systems like HeidelTime [58, 59] and offered by TINT as temporal tagging component.

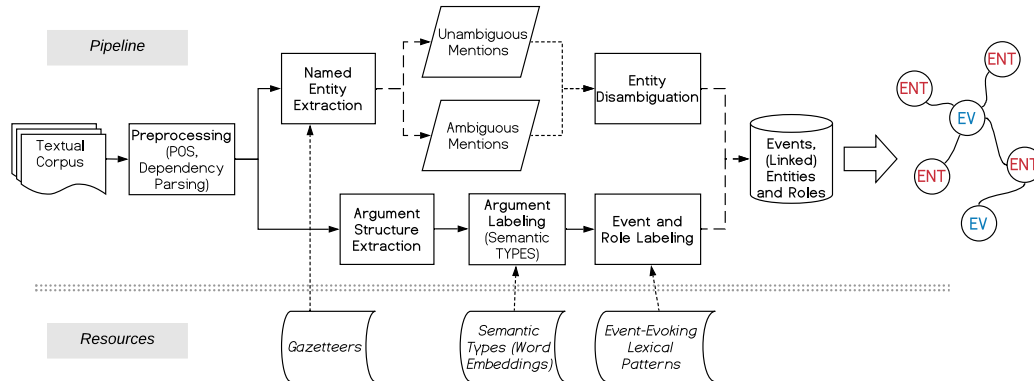


Fig. 2. The full text-to-knowledge pipeline described in this article.

5 DESCRIPTION OF THE SYSTEM

In this section the implementation of our system for Event and Participants Extraction is described. The pipeline (sketched in Figure 2) is composed of three macro-steps: (1) extraction and disambiguation of Named Entities, (2) tagging of lexical units' arguments with semantic types, (3) recognition of events and their participants. Tasks (1) and (2) are described in Section 5.1, while task (3) in Section 5.2. As general-purpose NLP pipeline, we use TINT [2], an open source NLP pipeline for Italian based on Stanford CoreNLP. This software is employed for all pre-processing tasks (tokenization, POS-tagging, dependency parsing, and NER).

5.1 Named Entity Extraction and Disambiguation

The first step of our pipeline consists of extracting all mentions of Named Entities (Persons, Locations, and Organizations) and linking them to the respective unambiguous entries in the gazetteer; to do so, we employ the following strategy:

5.1.1 Surface Form Recognition. For Locations and Organizations, we identify, for each entry and for each lexical variation listed in the gazetteer, the corresponding textual mentions. For entities of type Person the look-up process is slightly more complex, since persons in our collection are often referred to by different combinations of their name, surname, and nickname (if any). Therefore, we first pre-process the Person gazetteer and produce, for each entry, a list of lexical patterns by combining these three building elements¹⁴; then, for each Person entry, we identify in text all potential mentions.

5.1.2 Disambiguation and Linking. All the identified textual mentions that correspond to a unique entry in the respective domain gazetteer, i.e., all non-ambiguous mentions, are directly linked to the knowledge base. For all the remaining mentions, that is, for all those textual mentions that correspond to more than one entry in the gazetteers, a disambiguation step is required for them to be linked to the correct entry. While for entities of type Location and Organization ambiguity in our corpus is more or less negligible, for Persons it is highly relevant. In particular, the following two types of ambiguity need to be resolved: cross-classes and intra-classes.

Cross-classes ambiguity. In a certain number of cases, a given surface form can refer to entities, which may belong to more than one entity class. This is often the case for persons' surnames that are also toponyms (*Genova*, *Alessandria*, *Siracusa* are some examples) or for organizations that took their name from fallen fighters (*Brigata Rolando Besana*) or activists (*Brigata Carlo Rosselli*). Although the phenomenon is not pervasive—it affects around 1,000 instances in the Memoirs corpus—it is a feature of our textual corpus and it causes a systematic drop of

¹⁴For example: “Name Surname Nickname,” “Name Nickname Surname,” “Nickname Surname,” and so on. We identified 26 frequent patterns.

performance. Moreover, the issue often arises when a single-token surface form is used, like in the sentence *Ai reparti della Besana toccò schierarsi sulla destra del dispositivo*.¹⁵ To tackle this problem, we used the sentences annotated by the previous direct linking phase—given their high precision—to train a classifier that is able to assign the correct label (PER, LOC, ORG) to the multiclass ambiguous mentions. Four different classifiers have been tested: K-Nearest Neighbor, Nearest Centroid, Naive Bayes, Support Vector Machine. The set of features used for training is the following:

1. the word preceding the Named Entity (NE) occurrence;
- 2–3. the part of speech of the 2 words preceding the NE occurrence;
- 4–5. the part of speech of the 2 words following the NE occurrence;
6. the average word embedding¹⁶ of the 2 words before the NE occurrence;
7. the type of dependency linking the NE to the verb;
8. the type of verb linked to the NE (movement, social, conflict).

The learning algorithm achieving the best results in prediction was the k-NN (k = 9), scoring 0.75 F1-score (macro) on a sample of 250 manually annotated sentences. We also observed that features 2, 3, and 7 are most relevant for the task.

Intra-class ambiguity. To deal with intra-class ambiguity (which affects the most number of cases), we rely on the basic assumption from the entity linking literature [53, 62] that co-occurrence is a valuable source of information for determining the identity of an entity mention. We build an iterative process where each ambiguous mention is resolved by considering how frequently each of the possible candidates appears with the other (already disambiguated) surrounding mentioned entities in the same sentence. First, we compute an association score between each pair of disambiguated entities in the whole corpus (Memoirs) in the following way:

$$\text{association} = \frac{\text{co_occur}(A, B)}{\left(\frac{f_A + f_B}{2}\right)},$$

where $\text{co_occur}(A, B)$ represents the co-occurrence, in absolute terms, between the two entities A and B, while the denominator represents the averaged sum of their overall frequency.^{17, 18} Then, given an ambiguous mention, we rank each candidate for that mention based on the association score between the candidate and all the disambiguated entities appearing in context. The ambiguous mention is then linked to the candidate achieving the highest score. Let us consider the following example given the verb *rientrati* (they came back):

*19 marzo 1945 [...] Sono rientrati Renato, Saro, Nino, Pino, Marco, Carlin e Siracusa con 26 (sic) uomini, non cattivi, ma non dei migliori [...].*¹⁹

As a first step, the cross-class ambiguity of the argument *Siracusa* (both a city name and the nickname of a partisan) will be resolved using the previously discussed method (k-NN). Having recognized it as a named person present in our gazetteer, we will use this information to disambiguate the other 6 names considering the following numbers of candidates: *Renato* (15 candidates), *Saro* (2 candidates), *Nino* (25 candidates), *Pino* (8 candidates),

¹⁵“The units of the ‘Besana’ had to be deployed to the right.”

¹⁶We expand on the use of word embeddings in our work in the next sub-section.

¹⁷We also experimented using Pointwise Mutual Information for this task, achieving slightly worse results for the whole task.

¹⁸This measure is characterized by several features: It (a) is symmetric; (b) ranges from 0 (no association) to 1 (perfect association), the two entities always appear together in the corpus); (c) uses the average as denominator to “normalize” the score according to the magnitude of the two frequencies. This is useful when the frequencies have different size orders; and (d) it is biased towards low frequency entities, in that it tends to overestimate association when both entities have a low absolute frequency—this is a desirable feature in our use case, as we have a long tail of sparse entities, but if needed it can be circumvented by applying a cutoff threshold.

¹⁹“19th March 1945 [...] Renato, Saro, Nino, Pino, Marco, Carlin and Siracusa came back with 26 (sic) men, not bad ones, but not the best [...].”

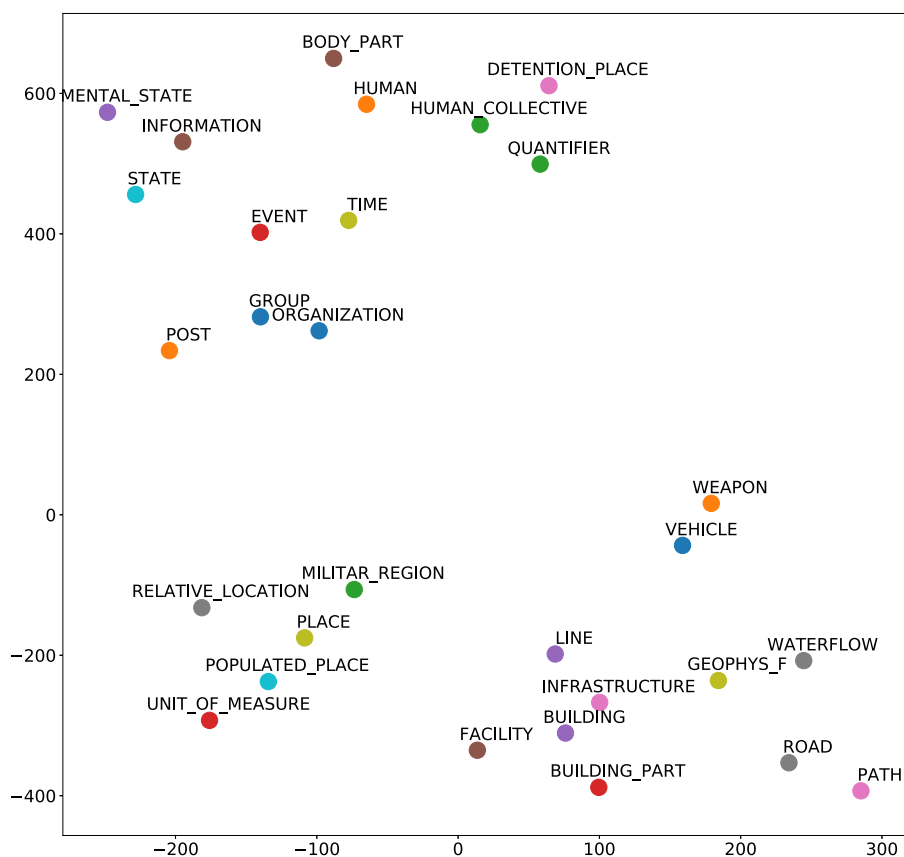


Fig. 3. A plot in two dimensions of the centroids representing semantic types (created with t-SNE).

Marco (14 candidates), *Carlin* (4 candidates). In the evaluation section, we provide evidence that addressing intra-class ambiguity this way permits us to identify over 3,000 additional disambiguated entity mentions.

For all the remaining named entities not listed in our gazetteers as well as for time expressions, we integrate in our system the output of the NER module of the TINT pipeline, which tags them as PER, LOC, and ORG.

5.2 Semantic Type Classification

NEs represent only a subset of the semantic types of lexical fillers in the argument structure of an LU. To identify the type of the other arguments related to a given syntactic head, we combine our dictionary of semantic types (see Section 3.2) with word embeddings [31], a relatively recent computational linguistic technology grounded on the distributional hypothesis [22]. We use 300 dimensional pre-trained fastText embeddings²⁰ [7] in the following way:

- (1) We start by creating a centroid for each semantic type in the dictionary (BUILDING, VEHICLE, HUMAN_COLLECTIVE, etc.) as the averaged sum of the word embedding vectors of the nominal words belonging to that type; this centroid represents the “center of mass” of each semantic type. A visual representation of the obtained centroids is provided in Figure 3.²¹

²⁰<https://dl.fbaipublicfiles.com/fasttext/vectors-crawl/cc.it.300.bin.gz>.

²¹The plot has been produced using t-SNE [27], perplexity=2, 5,000 iterations.

Table 4. Lexical Unit, Arguments, Syntactic Dependencies, and Semantic Types for the Sentence in the Example

Lexical unit:	<i>salire</i>	
Tagged args:	<i>ufficiali</i>	nsubj :: [HUMAN_COLLECTIVE]
	<i>Pinerolo</i>	nmod :: [a [LOC]]
	<i>Barge</i>	nmod :: [da [LOC]]
	<i>Valle Po</i>	nmod :: [in [LOC]]

- (2) In the same way, we represent each new argument to be tagged through the word embedding vector corresponding to its head.
- (3) Finally, arguments are tagged by computing the cosine similarity between their semantic vector and each of the centroids and assigning it to the closest one, as in a Rocchio classifier [51].

5.3 Event Classification and Role Labeling

Once arguments are labeled with semantic types, we have all elements in place to use our event-predicate dictionary. Let us consider the following example:

*Sempre negli stessi giorni dell'11 e 12 settembre, da Pinerolo salirono a Barge (in Valle Po) alcuni ufficiali di Cavalleria.*²²

The final goal of our pipeline is to classify the event triggered by the lexical unit *salirono* (they came up) according to the available event types and to assign a semantic role to each of the tagged arguments of the event anchor. To do so, we follow this procedure:

- (1) The sentence is represented as a set, consisting of the lexical unit and the annotated arguments, each argument labeled with its syntactic dependency, semantic type and preposition, as in Table 4.
- (2) Using our resource, all the possible event classes corresponding to the given lexical unit are retrieved. In this example, the LU *salire* can trigger three classes: BOARD_VEHICLE, MOVE_UPWARDS, and PATH_SHAPE;
- (3) Each event class is described in the event dictionary by a set of lexico-syntactic argument patterns; the set of tagged arguments from the sentence is compared with the sets of each candidate class and the class scoring the highest intersection with the tagged argument structure of the sentence is assigned to the LU. In the example, the result is the following:

MOVE_UPWARDS: 4
 PATH_SHAPE: 2
 BOARD_VEHICLE: 1

The sentence is thus assigned to the MOVE_UPWARDS type.

- (4) Finally, the semantic roles provided by the event class are assigned to each argument, which delivers a full-round semantic representation of the event mention, as presented in Table 5.

Two exceptions to our standard pipeline need to be discussed further: (a) often the given lexical unit maps only to one event class. In this case, we classify the event and assign roles if at least one tagged argument from the argument structure of the sentence matches the set of arguments provided by the event class (in other words,

²²“In the same days, 11th and 12th September, some cavalry officers came up to Barge (in the Po Valley) from Pinerolo.”

Table 5. Sentence Annotated with Event Type and Semantic Roles

Lexical unit:	<i>salire</i>		
Tagged args:	<i>ufficiali</i>	nsubj :: [H._COLLECT.]	MOVER
	<i>Pinerolo</i>	nmod :: [a [LOC]]	SOURCE
	<i>Barge</i>	nmod :: [da [LOC]]	GOAL
	<i>Valle Po</i>	nmod :: [in [LOC]]	GOAL
Event Class		MOVE_UPWARDS	

Table 6. Results of Event Extraction, Divided by Subcorpus, Confidence, and Type of LU Triggering the Event

Lexical type	Confidence	<i>Memoirs</i>	<i>Wiki</i>	<i>Biographies</i>
Verbals	High	3,583	3,151	270
	Low	11,823	8,190	454
Nominals	High	129	140	8
	Low	1,985	3,106	133
Multi-word verbal	High	109	202	12
	Low	503	679	55
	High	3,821	3,493	290
	Low	14,311	11,975	642
		18,132	15,468	932

if in the above mentioned procedure, it scores at least 1 in step 3). (b) In case of ties between two or more classes, we currently do not tag the event, as for our final application, we value precision over recall.²³

Since the system assigns a class label based on the type and number of the tagged arguments of the lexical unit, we suppose that the more arguments of an LU the system is able to tag correctly, the higher the probability is for the class label to be correct. This hypothesis will be substantially confirmed while evaluating the system (see Section 6.3). Therefore, in Table 6, we present the results by keeping high and low confidence results separated. Low confidence mentions are event mentions that have been extracted and classified by the system based on *one single* matching tagged argument, while high confidence mentions are based on at least two tagged arguments.

6 EVALUATION

In this section, we present a quantitative evaluation of each step of our pipeline.

6.1 Named Entity Extraction and Disambiguation

The strategy employed for extraction and disambiguation of Named Entities from our corpus consists basically of two steps: first, all non-ambiguous mentions are retrieved and linked through dictionary look-up (which we use as a baseline); then the remaining ambiguous mentions are disambiguated using the method described in Section 5.1.2. In this work, we deal with the problem, well known in the Digital Humanities and Digital Library communities, of disambiguating domain-specific entities, for which no entry in general purpose knowledge

²³The reason for this is twofold. First, in a rule-based system like ours, Recall performance is mostly due to the richness of the lexical pattern dictionary, which is by definition always extensible. At this stage, we aim to understand if and to which extent the patterns are productive. Second, since we envisage to use the data for visual representation/querying, we prefer to provide precise result, at the cost of incompleteness.

Table 7. Evaluation of Entity Disambiguation on a Gold Standard of 400 Manually Annotated Sentences (Precision-oriented Scenario)

PER		
Precision	0.842	(−0.056)
Recall	0.669	(+0.115)
F1 score	0.746	(+0.078)
Linked mentions	18,709	(+3306)
Ambiguous mentions	11,791	
LOC		
Precision	0.936	(+0.000)
Recall	0.927	(+0.007)
F1 score	0.931	(+0.003)
Linked mentions	26,761	(+257)
Ambiguous mentions	959	
ORG		
Precision	0.931	(+0.000)
Recall	0.911	(+0.000)
F1 score	0.921	(+0.000)
Linked mentions	4,943	(+34)
Ambiguous mentions	34	

In brackets, the increase/decrease of performance with respect to the baseline (direct linking of non-ambiguous entity mentions).

bases (e.g., DBpedia) is available [29, 35, 47]. We report the evaluation of the disambiguation strategy based on co-occurrence of entities in text and discuss its limitations. Table 7 shows the results of the evaluation of the system based on a gold standard of 400 sentences manually annotated with Named Entities and their link to the respective gazetteers.

For each entity type, we report Precision and Recall of the disambiguation/linking task, along with the absolute number of linked mentions and the remaining ambiguous mentions. For Locations and Organizations the performance is fairly high and the disambiguation step does not improve significantly the score on the baseline. This is due to the low intrinsic ambiguity of these entity types. Concerning Persons, in the chosen setting, our disambiguation strategy allows to improve the Recall of more than 11 points by losing around 5 points in Precision, leading to an overall increase of the F1-score of almost 8 points.

Error Analysis (Persons). Through the analysis of false positives and false negatives, we can dig deeper into the causes of such errors. Regarding Precision, it turns out that only 48% of errors are due to a wrong classification of the system, while the remaining 52% are represented by cases of homonymy either caused by entities not present in the gazetteers or by mismatches between Named Entities and other parts-of-speech. Where Recall is concerned, we observed that (a) more than 90% of false negatives occur with single-token mentions, mainly first names or surnames and, more importantly, that (b) the mentions are isolated, i.e., the context in terms of other disambiguated mentions is very poor. This situation prevents the ranking system to work, as all candidates automatically get a zero score. It is also important to note that most of the false negatives occur in texts adopting a diaristic style. As a whole, these observations show that the disambiguation strategy is effective as long as enough contextual information is provided.

Table 8. Evaluation of the Embedding-based Tagging System on a Gold Standard of 252 Sentences

	Correct	Wrong	Accuracy (%)
@1	190	62	0.754
@3	207	45	0.821

Table 9. Absolute Number of Anchors and Events Extracted from the Memoirs-test Set

LU type	Anchors	High	Low	Sum
verbs	2,002	145	574	719 (73.2%)
nouns	1,224	12	207	219 (22.3%)
multi-word verbs	87	6	38	44 (4.5%)
	3,313	163 (16.6%)	819 (83.4%)	982 (100%)

6.2 Semantic Type Classification

For evaluating the performance of the semantic tagging system described in Section 5.2, a test set of 252 words has been created by randomly picking argument fillers of verbs, nouns, and multi-word verbal expressions and by manually annotating them. For the evaluation, the TIME semantic type (i.e., centroid) has been added, as in the normal tagging procedure TINT it is used for recognizing temporal expressions. Moreover, an OTHER type has been used in the manual annotated gold standard for labeling words that do not belong to any of the available semantic types. On the automatic side, a threshold of 0.4 in the similarity score has been set, below which a word is tagged as OTHER. Results are shown in Table 8.²⁴

Error analysis. As shown in the table, the performance of the system is fairly good, but it does not improve significantly when the cut-off rank changes from @1 to @3. The reason for that can be further explained by looking at the incorrectly labeled examples from the test set. By focusing on the evaluation @3, it turns out that in 77% of the incorrectly labeled cases an OTHER annotation appears, either in the gold standard or in the automatic annotation; this fact suggests that the main source of error is given by the incompleteness of the adopted semantic type system, which only accounts for a limited set of entity types, the ones more relevant for the domain. For example, the adopted semantic types do not account for abstract entities (knowledge, feelings, intentions, etc.). The remaining 23% of incorrectly labeled words (10 cases) represents the genuine source of error, i.e., inter-class error.

6.3 Event Classification and Role Labeling

We approach the task of (a) extracting events and participants and (b) labeling participants with roles in different sub-steps. The evaluation of the two tasks is therefore performed separately and it is conducted on the Memoirs-test subcorpus, which was not employed for the creation of the event-predicate dictionary. This corpus has been processed with our pipeline, as described in Section 5; Table 9 summarizes the number of extracted and classified events.

²⁴Given the small size of the test set, we do not perform a tag-wise evaluation; instead, we report the overall accuracy (i.e., percentage of correctly labeled words).

Table 10. Evaluation of the Event Extraction Task on a Gold Standard of 300 Sentences

Precision	0.78
Recall	0.50
True negatives	0.88
F1-score	0.61

Extraction. For evaluating the extraction performance, we manually annotated 300 sentences randomly chosen from the test subcorpus, among the sentences that contain at least one lexical unit modeled in the event-predicate dictionary. The annotation is binary and assesses whether or not the sentence denotes an event mention. Results are summarized in Table 10. By analyzing the types of error, it turns out that the loss in Precision is mainly due to metaphoric use of language or non-compositionality of the predicate²⁵ (73% of cases) or to previous errors in the pipeline (27%). The low Recall has three main reasons: the lack of coverage of the resource, along with the misclassification with different lexical meanings (47% of cases); the presence of errors in previous tasks of the pipeline, especially in dependency parsing (17%); and the mention of events that have no relevant syntactic dependencies (17%), which prevents the system from assigning any class label to the word and to recognize the word as an event trigger at all (see step 3 in Section 5.3).

Classification. Given an event mention extracted by the system, we are interested to know how precise our pipeline is at assigning a class label to the mention, i.e., at classifying the event type. For evaluating this step, we randomly collected 200 event mentions extracted by the system and annotated them manually with the correct event class. The gold standard is further divided in two parts: 100 mentions are chosen among the “low confidence” mentions, while the remaining 100 are “high confidence” mentions. The results confirm the hypothesis formulated in Section 6.3, since the low confidence set leads to a 0.73 of Precision, while in the high confidence set the precision grows up to 0.89. Overall, these numbers also demonstrate that, when the system is provided with enough correctly extracted semantic information, it performs very well. It should also be considered that this last step is built on top of a long pipeline, thus suffering from a downstream propagation of errors in previous steps.

7 REPRESENTATION

The system described so far produced a knowledge base containing occurrences of events, entities (participants), and qualified relationships (roles) between entities and event mentions such entities participate in. To explore the results of the system and visualize the relationships between entity and event occurrences, we represent the extracted knowledge as a network of events, participants, and roles²⁶; the resulting graph has two applications:

- (1) providing a structure for querying the event knowledge base;
- (2) providing a basic visualization tool.

The network has two types of nodes: event nodes and entity nodes. Since Named Entities (as well as some time expressions) are linked, nodes representing the linked entities appear only once in the graph, while non-linkable entity types (HUMAN COLLECTIVES, VEHICLES, GEOPHYSICAL FEATURES, etc.) appear multiple times, once per mention. Entity nodes are also labeled with the type of the entity and with the corresponding text span. Event nodes are linked to entity nodes by an edge labeled with (a) the semantic role played by the entity in the event and (b) the document the event-entity pair has been extracted from.

²⁵Some examples: *respingere una proposta*, “reject a proposal,” tagged as REPEL; *incontrare resistenza*, “meet (with) resistance,” tagged as ENCOUNTER; *tornare a perdere il sentiero*, “lose the trail again,” tagged as RETURN.

²⁶A Gephi version of the event graph is available for download at <https://bit.ly/2UIZISo>.

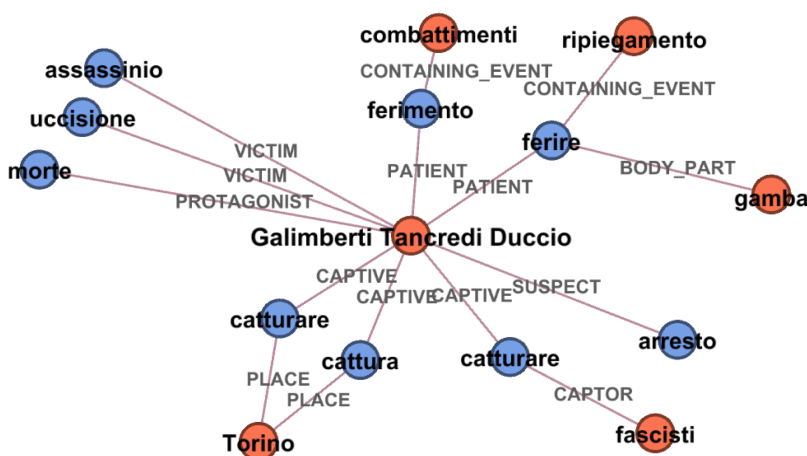


Fig. 4. Injuring, arrest, and assassination of “Duccio” Galimberti, extrapolated from the event graph. Nodes in blue represent events, nodes in orange represent entities. Edges are labeled with semantic roles.

This structure allows, given a named entity (or, better, given its node in the graph), to directly access all the events in which that entity participates. Moreover, given an event, it is always possible to go back to the document and the sentence where that event is mentioned.

If we consider the example in Figure 4, one of the events in which the central entity, *Tancredi “Duccio” Galimberti* has been involved, was his arrest (“cattura”—in the bottom left part of the graph). This event is modeled using information extracted from specific sentences in different books, to which we can always go back directly from the graph. For instance, we would see how he was captured by the fascists (in the first sentence) and how this event took place in Turin (second):

*Individuato un recapito delle staffette delle formazioni i fascisti vi avevano catturato Galimberti, mentre ritirava le lettere.*²⁷

*Un grosso colpo per noi in quel periodo fu la cattura di Duccio Galimberti a Torino, catturato davanti alla panetteria dove aveva il recapito clandestino.*²⁸

As highlighted below, this type of representation offers intermediate access to texts between Close and Distant Reading, by keeping each piece of knowledge (event or entity mention) within its more general context (the whole textual corpus, represented by the whole graph). By using this representation structure it is possible to benefit from bidirectional semantic access to texts, both from the text to the knowledge (events, entities, and roles) and the opposite way from any piece of knowledge to its textual source, with the only constraint being the design choices given by our event model (event classes and semantic roles).

7.1 Access Scenarios

What follows is a brief sketch of three possible access scenarios enabled by our representation structure.

Egocentric networks. An egocentric network (or ego network) is a graph centered on a specific node and it is used to highlight the relationships of the ego node²⁹ with other nodes. In our case, ego networks can be employed

²⁷“Having found an address for the formations’ couriers, the fascists had captured Galimberti there, while he was collecting the letters.”

²⁸“A big shock for us in that period was the capture of Duccio Galimberti in Turin, caught in front of the bakery where he had his clandestine address.”

²⁹An egocentric or ego node is the focal/central node in an egocentric network.

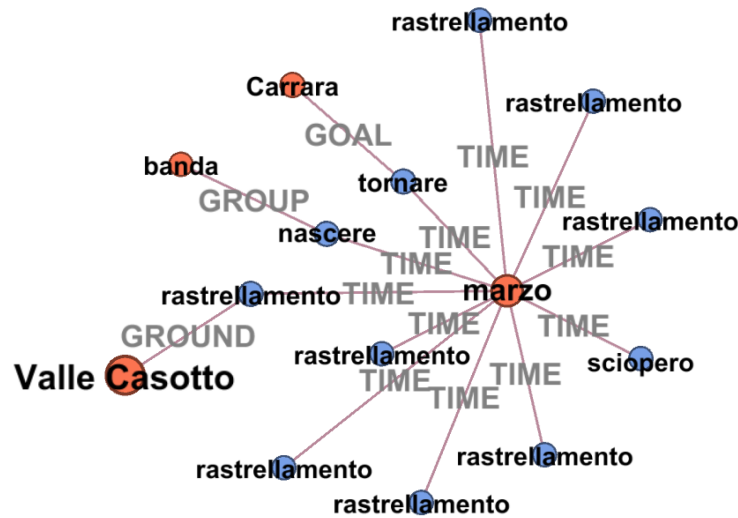


Fig. 5. Time-based egocentric network, depicting events happened in March 1944 (mentions thereof in the Memoirs corpus).

for entity-driven access by obtaining all events a given entity is involved in. If we return to Figure 4, this depicts all extracted events concerning the person of *Tancredi “Duccio” Galimberti*, a political leader of the Resistance movement. In the top part of the network, we can see the event of his injuring (mentioned twice in two different documents), and at the bottom is his capture (arrest) in Torino at the hands of fascist militias—mentioned 4 times in 4 documents—and left, his murder (3 mentions in 2 documents). Another way to use ego networks is to extract the subgraph related to a given date; for example, Figure 5 shows events happened in the month of March 1944. This access mode is particularly suited for users who already have a focus of interest and need to contextualize it within the dataset.

Constrained queries. Our representation structure can be further leveraged for building advanced queries according to different information needs. By introducing constraints in terms of semantic types and roles, we retrieve all entities that fulfill certain roles. For example, people undergoing an arrest can be retrieved by constraining entities on type “PER” and “HUMAN,” events on type “ARREST” and semantic roles on “SUSPECT.” Figure 6 shows a subset of arrest and capture events from the graph. Along this line, we can retrieve, for example, deported people, cities or places occupied or liberated or where clashes/battles took place, persons who created or joined organizations, and so forth. This scenario may be useful for users who are interested in a specific event type and want to examine its coverage in the dataset.

Events and entities as summarization tool. A third scenario supported by our system is the extraction of the subgraph of events and entities related to a specific document or a set of documents. This feature can be seen as a high-level summarization tool and provides the user with a way to compare quickly two documents.

8 REUSABILITY AND FUTURE STEPS

While our work has focused on a highly specific and challenging research domain, the event-extraction pipeline described above is *per se* domain agnostic and can be adapted to applications in many other scenarios; the domain-dependent component is represented by the event-predicate dictionary (and by semantic types), which should be updated or integrated according to the need. To do this, it could be sufficient to compare the set of our annotated lexical units with the target set of LU of interest in the new domain corpus to assess which LUs and event classes should be updated. The effort required in this step is difficult to predict, as it depends heavily on

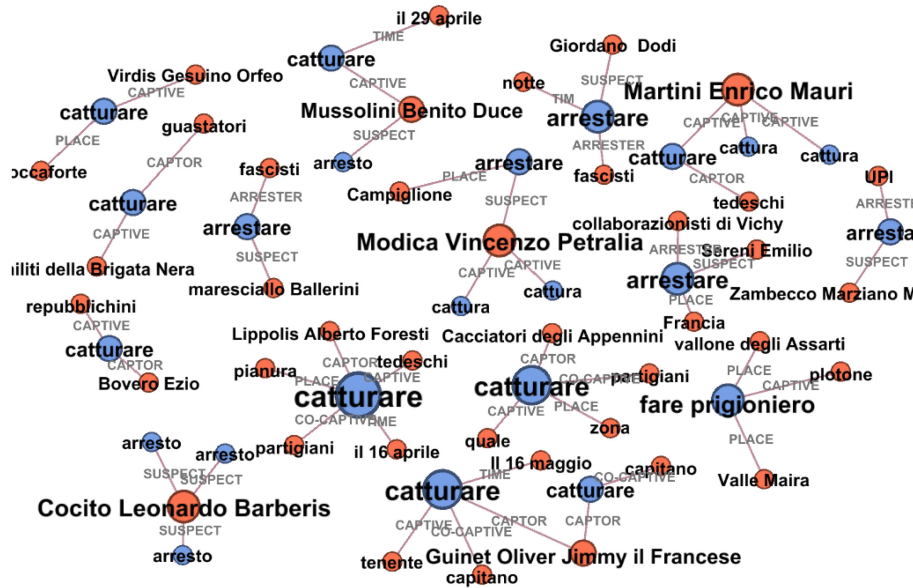


Fig. 6. Arrests and capture events from the graph (sample).

the new domain and on the extraction target; nevertheless, the existing literature on pattern harvesting has shown that in many cases a few patterns are sufficient to cover most occurrences of a given LU (thus drastically reducing the so-called “lexical entropy”) [21]. Moreover, by relying on our dictionary of event-predicate patterns, researchers could extract events, participants, and roles from Italian news articles, by combining our resource with the output of general-purpose entity linking systems (e.g., TagMe [16]), which we could not employ in our work due to the lack of domain coverage. Similarly, our gazetteer could be employed by researchers studying the same historical events, but employing primary sources written by German or American soldiers, due to the fact that the relations between people, places, and organizations will stay the same. Finally, this work could be seen as a set of guidelines for researchers aiming to build event-centered networks from historical sources, to whom we highly recommend to start by studying the coverage of resources such as FrameNet for their domain/language and, when lacking, the possibility of modeling the event types under study by extending our relational schema.

Based on the experience gained during this work and the results obtained through the network-based visualization, we now envisage three main research directions in the next future:

- (1) *Event-predicate lexical dictionary.* The event-predicate dictionary has proved to be a very useful and effective knowledge resource for extracting information from text. Since it partially uses FrameNet classes, it can directly be linked to this widely used resource in a multilingual setting. Our goal is to widen the resource in terms of coverage by both integrating the existing event classes and by taking into account new lexical units and event classes from other domains. Given the considerable manual effort for populating such a resource, a semi-automatic strategy must be devised for this purpose.
- (2) *Anaphoric expressions.* As shown in the evaluation (especially in Section 6.3), the presented system is very sensitive to the lack of information in the structure of the given lexical unit. However, in discourse such information is often “hidden” due to anaphoric use of language (which is rather pervasive in Italian). Therefore, being able to resolve anaphoric expressions is a key step to improve both recall and precision of the system, without changing the overall methodology.

- (3) *Event Coreference*. The described graph architecture provides a starting point for further work on Event Coreference, that is, the task of linking different (textual) event mentions to the corresponding (real) event occurrence. Figure 5 is a suitable example, as it represents nine event mentions that are actually referencing to three real event occurrences: the injuring, arrest/capture, and assassination of Duccio Galimberti. To this effect, our extraction pipeline could be used, alone or combined with other inputs, for producing a set of candidates for the coreference task. Being able to resolve automatically coreference between event mentions is our next step, as it would provide a very powerful tool for linking different information sources and to discover new information.

APPENDIX

A LIST OF DIGITIZED TEXTS

This section contains the bibliographical entries of the digitized books that constitute the Memoirs and the Memoirs-test corpora (see Section 3.1). Beside each entry, the corresponding abbreviation is specified. In square brackets, we report the year of first publication if it differs from the year of the edition used for OCR/digitization. Entries marked with a star (*) have been obtained by automatic text conversion from an existing digital edition.

Memoirs corpus:

- Aimo Renato, *Resistenza senza miti: dalla Stura alla Vesubie la brigata G.L. "Carlo Rosselli,"* Cuneo, L'Arciere, 1991.
- Artom Emanuele, *Diari di un partigiano ebreo: gennaio 1940 - febbraio 1944*, a cura di Guri Schwarz, Torino, Bollati Boringhieri, 2008 [1966].
- Balbo Adriano, *Quando inglesi arrivare noi tutti morti. Cronache di lotta partigiana: Langhe 1943-1945*, Torino, Blu, 2005.
- Berga Ugo, *Diario partigiano: dall'8 settembre 1943 alla liberazione, gli eventi e le persone che coinvolsero la 106. brigata Garibaldi Giordano Velino*, s.l., s.n., stampa 2003 (Almese : Morra).
- Bianco Alberto, *Alberto Bianco: testimonianza partigiana*, a cura di Michele Calandri, Alessandra Demicellis, Savigliano, L'artistica, 1999.
- * Bocca Giorgio, *Partigiani della montagna: vita delle divisioni Giustizia e libertà del Cuneese*, Milano Feltrinelli, 2005 [1945].
- Burdino Felice Luigi, *Diario partigiano*, Pinerolo, Alzani, 2005.
- Chioldi Pietro, *Banditi*, Torino, Einaudi, 1975 [1946].
- Comollo Gustavo, *Il commissario Pietro, ANPI Piemonte*, stampa 1979 (Savigliano : Nuove Arti Grafiche).
- Cordero Italo, *Ribelle: esperienze di vita partigiana dalla Val Casotto alle Langhe dalla Liguria alle colline torinesi*, Mondovì, Fracchia, 1991.
- Diena Marisa, *Guerriglia e autogoverno: Brigate Garibaldi nel Piemonte occidentale 1943-1945*, Parma, Guanda, 1970.
- Diena Marisa, *Un intenso impegno civile: ricordi autobiografici del Novecento*, Torino, Lupieri, Fondazione Istituto piemontese Antonio Gramsci, 2006.
- Donadei Mario, *Cronache partigiane: la banda di Valle Pesio*, Cuneo, L'Arciere, 1980 [1973].
- Dunchi Nardo, *Memorie partigiane*, Cuneo, L'arciera, 1982.
- Giaccardi Giovenale, *Le formazioni R nella lotta di liberazione*, Cuneo, L'arciera, 1993 [1980].
- Giovana Mario, *Storia di una formazione partigiana: Resistenza nel Cuneese*, Torino, Einaudi, 1964.
- * Marchesini Gobetti Ada, *Diario partigiano*, Torino, Einaudi, 2014 [1956].
- Martini Mauri Enrico, *Partigiani penne nere: Boves - Val Maudagna - Val Casotto - Le Langhe*, Torino, Edizioni del Capricorno, 2016 [1968].
- Modica Vincenzo, *Dalla Sicilia al Piemonte: storia di un comandante partigiano*, 2. ed. riv. e corr, Milano, FrancoAngeli, 2003.

- * Revelli Nuto, *Le due guerre: guerra fascista e guerra partigiana*, Torino, Einaudi, 2005 [2003].
- * Revelli Nuto, *La guerra dei poveri*, Torino, Einaudi, 2014 [1962].
- * Ronchi Della Rocca Icilio, *Ricordi di un partigiano: la Resistenza nel Braidese*, a cura di Livio Berardo, Milano, Angeli, 2009 [1965].
- Sacchetti Aldo, *Un romano tra i ribelli: da Duccio Galimberti a Piero Cosa*, Cuneo, L'Arciere, 1990.
- Trabucco Angela, *Partigiani in Val Chisone: 1943–1945*, Torre Pellice, Tip. subalpina, 1959.
- Visetti Aurelio Dante, *Un ribelle come tanti: intorno ad un diario partigiano*, Cuneo, L'arciera, 1993.

Memoirs-test corpus:

- Aimo Renato, *Il prezzo della pace: la gente bovesana e la Resistenza*, 1943-45, Cuneo, L'arciera, 1989.
- Carminati Masera Diana, *Langa partigiana '43–'45*, Boves, Araba Fenice, 2007 [1971].
- Bianco Dante Livio, *Guerra partigiana*, Torino, Einaudi, 2006 [1954].

ACKNOWLEDGMENTS

The authors would like to thank Senan Kiryakos and Stella Peyronel for proofreading the article.

REFERENCES

- [1] James Allan, Ron Papka, and Victor Lavrenko. 1998. On-line new event detection and tracking. In *Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 37–45.
- [2] Alessio Palmero Aprosio and Giovanni Moretti. 2016. Italy goes to Stanford: A collection of CoreNLP modules for Italian. *arXiv preprint arXiv:1609.06204* (2016).
- [3] Mariona Coll Ardanuy and Caroline Sporleder. 2014. Structure-based clustering of novels. In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL'14)*. 31–39.
- [4] Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of the 17th International Conference on Computational Linguistics-Volume 1*. Association for Computational Linguistics, 86–90.
- [5] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *J. Mach. Learn. Res.* 3, Jan. (2003), 993–1022.
- [6] Cameron Blevins. 2014. Space, nation, and the triumph of region: A view of the world from Houston. *J. Amer. Hist.* 101, 1 (2014), 122–147.
- [7] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Trans. Assoc. Comput. Ling.* 5 (2017), 135–146.
- [8] Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*. Association for Computational Linguistics, 789–797.
- [9] Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the 47th Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2*. Association for Computational Linguistics, 602–610.
- [10] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L. Boyd-Graber, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Proceedings of the Conference on Advances in Neural Information Processing Systems*. 288–296.
- [11] Gregory Crane. 2006. What do you do with a million books? *D-Lib Mag.* 12, 3 (2006), 1.
- [12] Agata Cybulska 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*.
- [13] Agata Cybulska and Piek Vossen. 2013. Semantic relations between events and their time, locations and participants for event coreference resolution. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*. 156–163.
- [14] Agata Cybulska and Piek Vossen. 2015. “Bag of events” approach to event coreference resolution. Supervised classification of event templates. *Int. J. Comput. Ling. Appl.* 6, 2 (2015), 11–27.
- [15] George R. Doddington, Alexis Mitchell, Mark A. Przybocki, Lance A. Ramshaw, Stephanie M. Strassel, and Ralph M. Weischedel. 2004. The automatic content extraction (ACE) program-tasks, data, and evaluation. In *Proceedings of the International Conference on Language Resources and Evaluation*.
- [16] Paolo Ferragina and Ugo Scaiella. 2010. TAGME: On-the-fly annotation of short text fragments (by Wikipedia entities). In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*. ACM, 1625–1628.
- [17] Gerhard Gossen, Elena Demidova, and Thomas Risse. 2017. Extracting event-centric document collections from large-scale Web archives. In *Proceedings of the International Conference on Theory and Practice of Digital Libraries*. Springer, 116–127.

- [18] Shawn Graham, Ian Milligan, and Scott Weingart. 2015. *Exploring Big Historical Data: The Historian's Macroscope*. World Scientific Publishing Company.
- [19] Derek Greene and James P. Cross. 2017. Exploring the political agenda of the European parliament using a dynamic topic modeling approach. *Polit. Anal.* 25, 1 (2017), 77–94.
- [20] Patrick Hanks. 2004. Corpus pattern analysis. In *Proceedings of the Euralex International Congress*, Vol. 1.
- [21] Patrick Hanks and James Pustejovsky. 2005. A pattern dictionary for natural language processing. *Rev. Française Ling. Appliq.* 10, 2 (2005), 63–82.
- [22] Zellig S. Harris. 1954. Distributional structure. *Word* 10, 2-3 (1954), 146–162.
- [23] Stefan Jänicke, Greta Franzini, Muhammad Faisal Cheema, and Gerik Scheuermann. 2015. On close and distant reading in digital humanities: A survey and future challenges. In *Proceedings of the Eurographics Conference on Visualization (EuroVis)-STARS'15*.
- [24] Tarek Kanan, Xuan Zhang, Mohamed Magdy, and Edward Fox. 2015. Big data text summarization for events: A problem based learning course. In *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries*. ACM, 87–90.
- [25] Micki Kaufman. 2015. “Everything on paper will be used against me”: Quantifying Kissinger. In *Proceedings of the Digital Humanities Conference*.
- [26] Giridhar Kumaran and James Allan. 2005. Using names and topics for new event detection. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 121–128.
- [27] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9, Nov. (2008), 2579–2605.
- [28] Gideon S. Mann, David Mimno, and Andrew McCallum. 2006. Bibliometric impact measures leveraging topic analysis. In *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries*. ACM, 65–74.
- [29] Katherine McDonough, Ludovic Moncla, and Matje van de Camp. 2019. Named entity recognition goes to old regime France: Geographic text analysis for early modern French corpora. *Int. J. Geog. Inf. Sci.* 33, 12 (2019), 2498–2522.
- [30] Stefano Menini, Rachele Sprugnoli, Giovanni Moretti, Enrico Bignotti, Sara Tonelli, and Bruno Lepri. 2017. RAMBLE ON: Tracing movements of popular historical figures. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. 77–80.
- [31] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the Conference on Advances in Neural Information Processing Systems*. 3111–3119.
- [32] Teruko Mitamura, 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*. 66–76.
- [33] Franco Moretti. 2003. Graphs, maps, trees. *New Left Rev.* 24 (2003), 67.
- [34] Franco Moretti. 2013. *Distant Reading*. Verso Books, London.
- [35] Gary Munnely and Séamus Lawless. 2018. Investigating entity linking in early English legal documents. In *Proceedings of the 18th ACM/IEEE Joint Conference on Digital Libraries*. 59–68.
- [36] Federico Nanni, Hiram Kümper, and Simone Paolo Ponzetto. 2016. Semi-supervised textual analysis and historical research helping each other: Some thoughts and observations. *Int. J. Human. Arts Comput.* 10, 1 (2016), 63–77.
- [37] Federico Nanni, Simone Paolo Ponzetto, and Laura Dietz. 2017. Building entity-centric event collections. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL'17)*. IEEE, 1–10.
- [38] Federico Nanni, Simone Paolo Ponzetto, and Laura Dietz. 2018. Toward comprehensive event collections. *Int. J. Dig. Lib.* 21, 2 (2018), 215–229.
- [39] Robert K. Nelson. 2010. *Mining the Dispatch* (2010). <https://dsl.richmond.edu/dispatch/>.
- [40] David Newman, Youn Noh, Edmund Talley, Sarvnaz Karimi, and Timothy Baldwin. 2010. Evaluating topic models for digital libraries. In *Proceedings of the 10th Joint Conference on Digital Libraries*. ACM, 215–224.
- [41] Alexander C. Nwala, Michele C. Weigle, Michael L. Nelson, Adam B. Ziegler, and Anastasia Aizman. 2017. Local memory project: Providing tools to build collections of stories for local events from local sources. In *Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries*. IEEE Press, 219–228.
- [42] Ekaterina Ovchinnikova, Laure Vieu, Alessandro Oltramari, Stefano Borgo, and Theodore Alexandrov. 2010. Data-driven and ontological analysis of FrameNet for natural language reasoning. In *Proceedings of the International Conference on Language Resources and Evaluation*.
- [43] Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computat. Ling.* 31, 1 (2005).
- [44] Miriam Posner. 2015. Humanities data: A necessary contradiction. *Miriam Posner's Blog* 25 (2015).
- [45] James Pustejovsky, Bob Ingria, Roser Sauri, Jose Castano, Jessica Littman, Rob Gaizauskas, Andrea Setzer, Graham Katz, and Inderjeet Mani. 2005. The specification language TimeML. *The Language of Time: A Reader*. Oxford Press, 545–557.
- [46] Kirk Roberts, Cosmin Adrian Bejan, and Sanda M. Harabagiu. 2010. Toponym disambiguation using events. In *Proceedings of the International Florida Artificial Intelligence Research Society Conference*, Vol. 10. 1.

- [47] Marco Rovera, Federico Nanni, Simone Paolo Ponzetto, and Anna Goy. 2017. *Domain-specific Named Entity Disambiguation in Historical Memoirs*. Universitätsbibliothek Mannheim.
- [48] Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. 2016. *FrameNet II: Extended Theory and Practice*. Institut für Deutsche Sprache, Bibliothek.
- [49] Roser Saurí, Lotus Goldberg, Marc Verhagen, and James Pustejovsky. 2009. Annotating events in English. TimeML annotation guidelines. In *Proceedings of the 4th International Workshop on Semantic Evaluation (SemEval'09)*.
- [50] Karin Kipper Schuler. 2005. *Verbnet: A Broad-coverage, Comprehensive Verb Lexicon*. Ph.D. Dissertation. University of Pennsylvania.
- [51] Hinrich Schütze, Christopher D. Manning, and Prabhakar Raghavan. 2008. *Introduction to Information Retrieval*. Vol. 39. Cambridge University Press.
- [52] Roxane Segers, Marieke Van Erp, Lourens Van Der Meij, Lora Aroyo, Jacco van Ossenbruggen, Guus Schreiber, Bob Wielinga, Johan Oomen, and Geertje Jacobs. 2011. Hacking history via event extraction. In *Proceedings of the 6th International Conference on Knowledge Capture*. ACM, 161–162.
- [53] Wei Shen, Jianyong Wang, and Jiawei Han. 2015. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Trans. Knowl. Data Eng.* 27, 2 (2015), 443–460.
- [54] William V Spanos. 2017. Against distant reading: Retrieving close reading in the interregnum. *symplokē* 25, 1-2 (2017), 247–260.
- [55] Andreas Spitz, Satya Almasian, and Michael Gertz. 2017. EVELIN: Exploration of event and entity links in implicit networks. In *Proceedings of the 26th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, 273–277.
- [56] Andreas Spitz and Michael Gertz. 2016. Terms over LOAD: Leveraging named entities for cross-document extraction and summarization of events. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 503–512.
- [57] Rachele Sprugnoli and Sara Tonelli. 2017. One, no one and one hundred thousand events: Defining and processing events in an interdisciplinary perspective. *Natural Language Engineering* 23, 4 (2017), 485–506.
- [58] Jannik Strötgen, Thomas Bögel, Julian Zell, Ayser Armiti, Tran Van Canh, and Michael Gertz. 2014. Extending HeidelTime for temporal expressions referring to historic dates. In *Proceedings of the International Conference on Language Resources and Evaluation*. Citeseer, 2390–2397.
- [59] Jannik Strötgen and Michael Gertz. 2013. Multilingual and cross-domain temporal tagging. *Language Resources and Evaluation* 47, 2 (2013), 269–298.
- [60] Piek Vossen, Rodrigo Agerri, Itziar Aldabe, Agata Cybulska, Marieke van Erp, Antske Fokkens, Egoitz Laparra, Anne-Lyse Minard, Alessio Palmero Aprosio, German Rigau, et al. 2016. NewsReader: Using knowledge resources in a cross-lingual reading machine to generate more knowledge from massive streams of news. *Knowledge-Based Systems* 110 (2016), 60–85.
- [61] Matthew Wilkens. 2013. The geographic imagination of Civil War-era American fiction. *American Literary History* 25, 4 (2013), 803–840.
- [62] Ian H Witten and David N Milne. 2008. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. (2008).
- [63] Hui Yang, Tat-Seng Chua, Shuguang Wang, and Chun-Keat Koh. 2003. Structured use of external knowledge for event-based open domain question answering. In *Proceedings of the 26th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 33–40.

Received September 2019; revised June 2020; accepted June 2020