Providing Advanced Access to Historical War Memoirs Through the Identification of Events, Participants and Roles

Marco Rovera
Dipartimento di Informatica
Università di Torino
Italy
rovera@di.unito.it

Federico Nanni, Simone Paolo Ponzetto
Data and Web Science Group
University of Mannheim
Germany
federico,simone@informatik.uni-mannheim.de

ABSTRACT

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

KEYWORDS

Event Extraction, Entity Linking, Digital Humanities, Semantic Role Labeling, Distant Reading, Second World War

1 INTRODUCTION

The growing interest of cultural institutions for digitization of archival documents and resources, along with the availability of new, born-digital textual materials about historical topics, raise the question of how to provide to the users an account of the knowledge contained in such collections. From a computational point of view, the challenge 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 knowledge into knowledge that can be easily accessed, queried and visualized by end users.

During the last two decades, as applications of the general framework of Distant Reading in historical research [16, 29], several interesting 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 [23], Wilkens [51]). However, while 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. [21], Nanni et al. [30]). Additionally, they generally lack the possibility of rendering back semantic information in a more fine-grained way (e.g., by retrieving highly relevant sentences).

To face this issue and foster the adoption of advanced Natural Language Processing (NLP) technologies for providing semantically-enriched access to historical collections, in this paper we present a system that models a central component of historical scholarships, namely the concept of "event". To do so, we a) first identify the mentions of events and event participants in text, along with the semantic roles of participants, and b) then use them as conceptual pivots for extracting structured semantic information. We apply our methodology in a use case of Event and Participant Extraction on a corpus of newly digitized Italian war memoirs concerning the Second World War. We present an extended evaluation and error analysis of each step of our pipeline and we finally discuss how, by modeling events, entities and semantic roles, our system is able to provide advanced access through both distant and close reading of the collection under study (as in Bonfiglioli and Nanni [8]).

Outline. The rest of the paper is structured as follows: in Section 2 we discuss some of the most relevant previous works related to our study from the fields of Event and Entity Extraction and Knowledge Representation. Section 3 presents the way we model events in our work, which led to the creation of a lexical resource for extraction of events and participants. Section 4 describes the textual corpus and the knowledge resources used in our system. 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 conclude by presenting a solution for exploring our event-based knowledge graph; this allows moving from the textual collection to a network representing events, participants and roles, back again to the text.

2 RELATED WORK

In this section, we first offer an overview of related work 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 last fifteen years LDA topic modeling [5, 34] 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 for offering advanced access to scientific collections [26], to proceedings of political debates [17] and to historical corpora [33]. To face the limitations of LDA topics [11] and in order to extract information that is 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 [23] relies on these approaches for examining the Digital National Security Archive (DNSA) Kissinger Collections, while Menini et al. [27] uses 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 paper 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 a specific semantic role given the contextual event under study and highlight their network of relations.

Event Extraction. In this work, we rely on the use of event mentions as central pivots for providing advanced access to historical collections. We do so, following the large recent interest from the digital library community in event-based collection building [15, 22, 31, 32, 35]. In NLP and related areas, the notion of event has been modeled and applied to different subtasks, such as Question Answering [52], Topic Detection and Tracking [1, 24], Narrative Chain Induction [9, 10], Entity Disambiguation [38], Information Extraction and Retrieval [45, 46]. An interesting research thread has been carried out at VU Amsterdam, focusing on different aspects of events in historical texts and newswires, for example Extraction [12, 43] and Coreference Resolution [13, 14]. This thread was also part of NewsReader [50], a multilingual, EU-funded research project focused on Event and Information Extraction from newswires. In the interdisciplinary frame of events and historical research, [47] provides an overview on the evolution of the notion of event and the related techniques in the NLP community and tries to bridge the gap between these and the notion of event in the history research community. Our work builds upon this previous study by providing a real 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 automatic 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 on 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 explicit, SRL is generally approached as a supervised task. Typical lexical knowledge bases for SRL in English are FrameNet [4], VerbNet [41] and Propbank [37]. Since no such knowledge base exists for Italian, in this work we have created a lightweight resource, focusing on three high-level event categories: movements, conflictual events, and social events concerning membership and organizations.

3 MODELING EVENTS

Events, along with their participants, are the conceptual notion that we model in order to recognize their mentions in text. Given the immense relevance that events have in the historical domain [47], in this work we decided to approach three different types, namely spatial events (movements), conflictual events and membership in organizations. When referring to "participants", we intend not just humans or artifacts, but any entity that could have a role in an event, including spatial and temporal entities. The setting is inspired by FrameNet (FN),\(^1\) with the notable difference that we are only interested in modeling event information (Events and States), while many FN Frames also model other type of information.\(^2\)

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 Lexical Unit - 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 a LU, at least for verbal and multiword verbal expressions, is the main source of information about the participants of an event denoted by that Lexical Unit. It follows that there is a correspondence between the fillers of certain syntactic dependencies of a 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 words 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 [40], in this work we reduce the focus only to verbs, nouns and multiword verbal expressions. This choice is motivated by the evidence that these categories of words do 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. Like in FrameNet, each LU can map to one or more event types.

2. Syntactic Dependencies. For events denoted by verbs and by multiword verbal expressions we consider the active/passive subject (nsubj, nsubj:pass types in Universal Dependencies\(^3\)), 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

\(^1\) Event Classes discussed at point 5 are, to some extent, equivalent to FN Frames.
\(^2\) From an ontological point of view, Frames describe "types of situations" [36].
\(^3\) http://universaldependencies.org/it/dep/
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 type of the entity that appears as filler of a given syntactic dependency. While some types are more general (HUMAN, HUMAN_COLLECTIVE, POPULATEDPLACE), 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. By means of these three components (Lexical Unit, syntactic dependency, Semantic Type) it is possible to provide syntactic-semantic patterns associated to a LU (see Table 1 for an example).

Table 1: Example of syntactic-semantic argument patterns associated to the Lexical Unit “bombardare” (to bomb).

| Lexical Unit | Event Class | Syn-Sem pattern | Semantic Role |
|--------------|-------------|-----------------|--------------|
| bombardare   | DEPARTING   | nsubj :: PER    | Mover        |
|              |             | dobj :: LOC     | Source       |
|              |             | nsubj :: ORGANIZAT. | Mover |
|              | QUIT_GROUP  | nsubj :: PER    | Member       |
|              |             | dobj :: ORGANIZAT. | Group |

Table 2: Excerpts from the mapping of two lexical units to the correspondent Event Classes. In each event class, all syntactic-semantic argument patterns associated to the verb are associated to a Semantic Role.

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 in order 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 4.2.3) of which 52 correspond to a FrameNet frame, while 36 are specific for modeling events present in this collection. The latter concern in particular a) events that are not modeled by any FrameNet frame, for instance specific war-related Event Classes like DEPORTATION, LIBERATION or AIRDROP as well as fine-grained movements, like 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 a) to assess which type of event is mentioned in a given context, b) to associate a set of entities (Named Entities or other types of entities) to that event and c) to assess which role is played, with respect to the event, by each event participant (example in Table 2).

4 MATERIALS

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

---

4 The set of Semantic Types considered in this work is depicted in Figure 2.
5 PER, 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 HeideTime [46, 49] and offered by TINT as temporal tagging component.
4.1 Textual Collections

The collection employed in this work is divided into three subcorpora, as presented below. 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 system will be tested directly on it. 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 events.

**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 depicted facts goes from the 8th of September 1943 to the 25th 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 documents 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. This acquisition effort resulted in a textual corpus of ≈1.5 million words and over 95,000 sentences.

**Biographies.** The second subcorpus has been obtained by web scraping the Wikisource page dedicated to “Men and Women of the Italian Resistance”, 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 also appearing 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”. This large subcorpus counts ≈1.3 million words and 53k sentences. The category groups together all the entries connected to resistance movements against nazifascism during the 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, persons).

**Memoirs-test.** In order to evaluate the system and the lexical pattern dictionary on the task of Event Extraction, a test corpus has been created employing memoirs that we have initially excluded from the Memoirs collection. They are composed by 112,000 words and 5,100 sentences.

| Documents | Words | Sentences | Acquisition |
|-----------|-------|-----------|-------------|
| Memoirs   | 25 books | 1,469,000 | Digitization |
| Bio       | 189 entries | 57,400 | Web scraping |
| Wiki      | 1,748 articles | 1,364,000 | Web scraping |
| Mem-test  | 3 books | 112,000 | Digitization |

4.2 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 subcorpus 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 text (with their related abbreviations and acronyms). As we have already shown in a previous work [39], 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. 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 a human being like “signora”, “uomo” or “ragazza”, but also words referring to social roles like “capitano”, “madre”, “attivista”, “carceriere”, which denote human beings through the role they have in a social context. This small lexical resource is composed by 424 words and is employed in the semantic tagging step, that will be discussed Section 5.

**A dictionary of event-evoking lexical patterns.** The most relevant knowledge resource for the aim of this work is the dictionary of event-evoking lexical patterns, exemplified in Table 2. The main purpose of the lexical 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 [18, 19], which aims at assigning a sense to a Lexical Unit (verbs or nouns) 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 Lexical Units (word occurrences) to a concept (events, in our case) and to retrieve
the structure of such concept in text. As opposed to CPA, a) our pattern dictionary is not meant to map lexical senses but event mentions denoted by lexical units, 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 verb belong to (or denote) different Event Classes. For example, the pattern abbandonare\(^{10}\) :: 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 pattern dictionary counts 246 LUs, mapped to 88 Event Classes: 124 verbs (like “aderire”, “imprigionare”, “partire”\(^{11}\)), 77 nouns (e.g., “arresto”, “liberazione”, “arrivo”\(^{12}\)), 45 multiword verbal expressions (“aprire il fuoco”, “fare prigioniero”\(^{13}\)). 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.

5 DESCRIPTION OF THE SYSTEM

In this section the implementation of our system for Event and Participants Extraction is described. The system uses the resources previously presented and has been employed on the Memoirs sub-corpus. The pipeline is composed by 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 link 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 elements;\(^{14}\) 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 is more or less negligible, for Persons it is highly relevant. In particular, the following two types of ambiguity needs to be resolved: a) cross- and b) intra-classes.

Cross-class ambiguity. As far as ambiguity across different entity types is concerned, for instance persons’ surnames that are also toponyms (“Genova”, “Alessandria”, “Siracusa”) or for organizations which took their name from fallen fighters (e.g. “Rolando Besana Brigade”), we employ a k-Nearest Neighbours (k-NN) classifier that assigns the correct label (PER, LOC, ORG) to the multiclass ambiguous mention by employing the following features:

1. the word preceding the Named Entity (NE);
2-3. the part of speech of the 2 words preceding the NE;
4-5. the part of speech of the 2 words following the NE;
6. the average word embedding\(^{15}\) of the 2 words before the NE;
7. the type of dependency linking the NE to the verb;
8. the type of verb linked to the NE (movement, social, conflictual).

This approach was used to annotate 1,000 cross-class ambiguous mentions. The learning algorithm achieving the best results on the training set was the k-NN (k = 9), scoring 0.75 F1-score (macro). We observed that features 2, 3 and 7 are the most useful.

Intra-class ambiguity. In order to deal with intra class ambiguity, we rely on the basic assumption from the entity-linking literature [44] 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 (i.e., that could be linked to more than one entry of the same type) is resolved by considering how frequently each of the possible candidates appears with the other (already disambiguated) mentioned entities in the same sentence. First, we compute the strength of the co-occurrence between each pair of disambiguated entities in the whole corpus (Memoirs) in the following way:

\[
\text{strength} = \frac{\text{co}_\text{occur}(A, B)}{\sqrt{\text{f}_A \cdot \text{f}_B}}
\]

where co\text{occur}(A, B) the absolute frequencies of the considered pair of entities A and B appearing in context (the same sentence). The numerator represents therefore the absolute number of co-occurrences between the two entities, while the denominator represents the averaged sum of their overall frequency. Then, given an ambiguous mention, we rank each candidate for that mention

\(^{10}\) “to leave”.
\(^{11}\) “to join” (an organization), “to imprison”, “to depart”.
\(^{12}\) “arrest”, “liberation/release”, “arrival”.
\(^{13}\) “to open fire”, “to take prisoner”.

\(^{14}\) For example: “Name Surname Nickname”, “Name Nickname Surname”, “Nickname Surname”, etc. We identified 26 frequent patterns.

\(^{15}\) We expand on the use of word embeddings in our work in the next sub-section.
based on the strength between the candidate and all the disan-
bigated entities appearing in context. The ambiguous mention:
then linked to the candidate achieving the highest score.
If we consider the following examples having the event-evoking
verb "sono 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 [...]».

As a first step, the cross-class ambiguity of the argument "Sir-
cusa" (both a city and the nickname of a partisan) will be resolve
using the previously presented k-NN. Having recognized him as
named person present in our gazetteer, we will use this infor-
mation to disambiguate the other 6 names considering the followin
g numbers of candidates: Renato (15 candidates), Saro (2 candidates)
Nino (25 candidates), Pino (8 candidates), Marco (14 candidates
Carlin (4 candidates). In the evaluation section we provide evidenc
that addressing intra class ambiguity permits us to identify over
3,000 additional disambiguated entity mentions.
For all the remaining named entities that are not listed in ou
gazetteers as well as for time expressions, we integrate in our syster
the output of the NER module of the TINT pipeline, which tag
them as PER, LOC and ORG.

5.2 Semantic Type Classification

NEs represent only a subset of the semantic types of the lexica,
fillers in the argument structure of a lexical unit. To identify the type
of the other arguments related to an event anchor, we combine our
Semantic Type dictionary (see 4.2.2) with word embeddings [28],
a relatively recent computational linguistic technology grounded
on the distributional hypothesis [20]. We use 300 dimensional pre-
trained fastText embeddings17 [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 nom-
inar 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 Fig. 2.18
(2) In the same way, we represent each new argument to be tagged
by averaging its word embedding vectors.
(3) Finally, arguments are tagged by computing the cosine similar-
ity between their semantic vector and each of the centroids and
assigning it to the closest one, as in a Rocchio classifier [42].

5.3 Event classification and Role Labeling

Once arguments are labeled with Semantic Types, we have all
elements in place to use our event-evoking lexical pattern dictionary.
Let us consider the following example:

16 19th March 1945 [...] Renato, Saro, Nino, Pino, Marco, Carlin and Siracusa came
back with 26 (sic) men, not bad ones, but not the best [...]».
17 https://fasttext.cc
18 The plot has been produced using t-SNE [25], perplexity=2, 5000 iterations.

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

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 annotated with its
syntactic dependency, Semantic Type and preposition (if any).

| Lexical unit:  | salire |
|----------------|--------|
| Tagged args:   | ufficiali |
| renmod :: [a [LOC]] |
| rmod :: [da [LOC]] |
| rmod :: [in [LOC]] |

(2) Using our resource, all the possible Event Classes correspond-
ing 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 pattern dictionary by a
set of lexico-syntactic argument patterns; the set of tagged
arguments from the sentence is compared with the sets of each

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

Finally, the Semantic Roles provided by the Event Class are assigned to each argument, which delivers a full-round semantic representation of the event under study, as presented in the following result:

```
Lexical unit: salire
Tagged args: ufficiali
H._COLLECT.
Mover
Pinerolo
LOC
Source
Barge
LOC
Goal
Valle Po
LOC
Goal
```

Event Class | MOVE_UPWARDS
---|---

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, 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.

| Lexical type | Confidence | Memoirs | Wiki | Biographies |
|--------------|------------|---------|------|-------------|
| Verbals      | High       | 3583    | 3151 | 270         |
|              | Low        | 11823   | 8190 | 454         |
| Nominals     | High       | 129     | 140  | 8           |
|              | Low        | 1985    | 3106 | 133         |
| Mw verb expr | High       | 109     | 202  | 12          |
|              | Low        | 503     | 679  | 55          |
|              | High       | 3821    | 3493 | 290         |
|              | Low        | 14311   | 11975| 642         |

Table 4: Results of event extraction, divided by subcorpus, confidence and type of LU triggering the event.

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 a LU the system is able to tag correctly, the higher the probability is for the class label to be correct. This hypothesis will be confirmed while evaluating the system (see Section 6.3, Q2). Therefore, in Table 4 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 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 is basically two-step: first, all non-ambiguous mentions are retrieved and linked through dictionary look-up (this represents our 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, increasingly more relevant in the Digital Humanities and Digital Library communities, of disambiguating domain-specific entities, for which no entry in general purpose knowledge bases (e.g., DBpedia) is available [39]. We report the evaluation of the disambiguation strategy based on co-occurrence of entities in text and discuss its limitations. Table 5 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.

| Lexical Unit | Precision | Recall | F1 score | Linked mentions | Ambiguous mentions |
|--------------|-----------|--------|----------|-----------------|-------------------|
| LOC          | 0.936     | 0.927  | 0.931    | 26.761          | 959               |
| ORG          | 0.931     | 0.911  | 0.921    | 4943            | 34                |

Table 5: Evaluation of Entity Disambiguation on a gold standard of 400 manually annotated sentences (precision-oriented scenario).

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. In brackets, the increase/decrease of performance is presented, with respect to the baseline (the linking of non ambiguous entity 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 reasons 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% is 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.

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 multiword verbal expressions and by manually annotating them. The tagging system, as well as the list of semantic types, is not meant to be cross-domain and is not expected to provide good generalization outside the domain; therefore this evaluation procedure allows the creation of a fair test set, made of words that, at least in part, were not part of the centroids but still related to the domain. For the evaluation the TIME semantic type (i.e. centroid) has been added, as in the normal tagging procedure TINT is used for the extraction of 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 6.

| Correct | Wrong | Recall |
|---------|-------|--------|
| @1      | 190   | 62     | 0.754 |
| @3      | 207   | 45     | 0.821 |

Table 6: Evaluation of the embedding-based tagging system on a gold standard of 253 sentences.

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 wrong labeled examples from the test set. By focusing on the evaluation @3, it turns out that in 77% of the wrong 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 wrong labeled words (10 cases) represents the genuine source of error, that is, 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 this tasks aims at answering the two main questions presented below, which cover respectively the extraction and classification aspects. This is conducted on the Memoirs-test subcorpus (see Section 4.1), which was not employed in any of the steps that led to the creation of the event-evoking pattern dictionary. This corpus has been processed with our pipeline, as described in Section 5; Table 7 summarizes the number of events extracted from it.

| LU type | Anchors | High | Low | Sum |
|---------|---------|------|-----|-----|
| verbs   | 2002    | 145  | 574 | 719 (73.2%) |
| nouns   | 1224    | 12   | 207 | 219 (22.3%) |
| m-w verb| 87      | 6    | 38  | 44 (4.5%)   |

Table 7: Absolute number of anchors and events extracted from the Memoirs-test set.

Q1: How does the system perform at extracting events? For answering this question, 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-evoking dictionary. The annotation is binary and assesses whether or not the sentence denotes an event mention (of one of the types modeled in the resource). The results, summarized in Table 8, reveal the capacity of our pipeline for extracting events, from Lexical Units.

|                  | Precision | Recall | True negatives | F1-score |
|------------------|-----------|--------|----------------|----------|
|                   | 0.78      | 0.50   | 0.88           | 0.61     |

Table 8: Evaluation of the event extraction task on a gold standard of 300 sentences.

By analyzing the types of error, it turns out that the loss in Precision is mainly due to metaphoric use of language (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 though to recognize the word as an event trigger at all (see step 3 in Section 5.3).

Q2: How does the system perform at classifying the extracted events into event classes? 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, at least in terms of Precision, when the system is provided with enough correctly extracted semantic information, it shows very good performances.

7 REPRESENTATION AND FUTURE STEPS

In order to explore the results of the system and to visualize the relationships between entities and events, we built a graph with two types of nodes: events and entities. Since Named Entities (as well as some time expressions) have been 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, one time per mention. 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 mention has been extracted from. These two features are very useful since they allow, given a named entity (or, better, given its node in the graph), to immediately visualize all the extracted events in which that entity participates. Moreover, given an event, it is always possible to go back to the document, to the precise sentence where that event is mentioned.

Using this representation, it is possible to provide the user with powerful semantic searches, only constrained by the design choices given by our event model (classes and semantic roles). For instance, it is possible to create "ego graphs" centered on an entity of interest or select only certain specific types of events or, still, constrain the query to retrieve all entities that fulfill a certain event role (e.g., all arrested people are entities of type PER or HUMAN, involved in an ARREST event and linked to it by a SUSPECT edge). Another possibility is to obtain all events and entities mentioned in a given document or set of documents, employing our system as a semantic summarization tool. Moreover, 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. Figure 3 is a suitable example, as it represents nine event mentions which are actually referencing to three real events: the injuring, arrest/capture (bottom) and assassination (top left) of Duccio Galimberti, one of the political leaders of the Resistance movement in Northern Italy.

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

1) Event-evoking lexical dictionary. The event-evoking 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 showed in the evaluation (especially in Section 6.3, Q2), the presented system is very sensitive to the lack of information in the structure of the given Lexical Unit. On the other side, 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 in order to improve both recall and precision of the system, without changing the overall methodology.

3) Event Coreference. The graph structure described in Section 7 is a good starting point for further analysis of event structure and event similarity. Being able to automatically resolve 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.

8 CONCLUSIONS

In this paper we presented a methodology for extracting semantic knowledge from historical texts based on events, participants and roles. The proposed methodology has been applied to a practical use case employing a corpus of memoirs of Italian partisans of the Second World War; all the main steps of the process have been quantitatively evaluated. The notion of event, coupled with linked entities, proved to be a valuable conceptual pivot both in the extraction phase and in the subsequent knowledge aggregation and visualization steps, opening up exciting new ways for both distant and close reading of digitized collections, at the intersection of historical research, digital libraries and NLP.
