Domain-specific Named Entity Disambiguation in Historical Memoirs

Marco Rovera[1], Federico Nanni[2], Simone Paolo Ponzetto[2], Anna Goy[1]
[1]Dipartimento di Informatica, Università di Torino, Italy
{rovera,goy}@di.unito.it
[2]Data and Web Science Group, University of Mannheim, Germany
{federico,simone}@informatik.uni-mannheim.de

Abstract

English. This paper presents the results of the extraction of named entities from a collection of historical memoirs about the italian Resistance during the World War II. The methodology followed for the extraction and disambiguation task will be discussed, as well as its evaluation. For the semantic annotations of the dataset, we have developed a pipeline based on established practices for extracting and disambiguating Named Entities. This has been necessary, considering the poor performances of out-of-the-box Named Entity Recognition and Disambiguation (NERD) tools tested in the initial phase of this work.

Italiano. *Questo articolo presenta l'attività di estrazione di entità nominate realizzata su una collezione di memorie relative al periodo della Resistenza italiana nella Seconda Guerra Mondiale. Verrà discussa la metodologia sviluppata per il processo di estrazione e disambiguazione delle entità nominate, nonché la sua valutazione. L'implementazione di una metodologia di estrazione e disambiguazione basata su lookup si è resa necessaria in considerazione delle scarse prestazioni dei sistemi di Named Entity Recognition and Disambiguation (NERD), come si evince dalla discussione nella prima parte di questo lavoro.*

1 Introduction and Motivation

Current NLP techniques allow us to treat some types of historical textual resources provided by, among others, historical archives and libraries, as a source of information (and, in prospect, of knowledge) for automatic systems. Besides encyclopedic resources, libraries and archives provide many different types of texts, often spanning very specific geographical, individual or thematic contexts, for which current knowledge extraction systems may lack the suitable information. Nevertheless, the tasks of extracting, disambiguating and linking information provided by historical textual documents with respect to external knowledge bases is still a crucial step towards automatic access to written resources and for further employ of such knowledge in end-user applications (e.g. navigation, rich semantic search, creation of narrative chains). In order to address longer term tasks, such as event extraction from historical texts (Goy et al., 2015), we first addressed the task of extracting and disambiguating Named Entities (Persons, Locations and Organizations) from a corpus of historical memories of the "Liberation War" in Italy, during the Second World War. Due to the specificity of the domain and of the involved entities, state-of-the-art tools for Named Entity Recognition and Disambiguation show low performances, thus suggesting us to try to achieve our goal using a different approach. In this paper we present a collection of documents created by digitizing historical memoirs, together with an overview of the methodology we followed for the extraction and disambiguation of Persons, Locations and Organizations, as well as the results of the evaluation of its output in comparison with the output of two state-of-the-art systems. The outline of the paper is the following: in Section 2 some related projects are discussed, while in Section 3 the dataset used in the experiment is presented. Section 4 describes the test of two automatic NER tools (4.1) and the methodology devised for our experiment (4.2). In Section 5 the results of the evaluation are discussed, while Section 6 concludes the paper and outlines the next developments of the project.

2 Related Work

The work described in this paper is mainly related to Named Entity Recognition and Disambiguation (NERD) techniques and their application in the field of Digital Humanities (DH), in particular on historical texts. While NER refers to the task of identifying named entities in text and classifying them according to a set of categories, a Named Entity Disambiguation (NED) task is aimed at assigning a correspondence between an ambiguous surface form and the individual entity it refers to. Although analytically they can be considered as two separate tasks, the current availability of large, publicly accessible knowledge bases allowed to merge them into the task of Entity Linking (EL), which aims at linking a surface form from a text to the corresponding entry in a resource like DBpedia or Wikipedia (Barrière, 2016). A recent application of EL techniques in a DH context is presented in Brando et al. (2016), where the authors use a graph-based approach and exploit Linked Data for linking mentions of writers in a corpus of French literary criticism and scientific essays. Discussions and experiments on the use of third-party NER services on historical OCRed texts (typewritten memoirs of Holocaust survivors and old newspapers respectively) are provided by Rodriquez et al. (2012) and by Ehrmann et al. (2016), offering a starting point for our work, since they quantify, showing their limitations, the performances of NER such tools on specific historical texts (as also remarked in Nanni et al. (2017)). Also in the Italian DH research community, the interest for mining historical texts became more evident in the last years and leading to several interesting works. In Boschetti et al. (2014), for example, the authors describe the ongoing work of applying a full Information Extraction pipeline (from OCR digitization to data visualization) to war bulletins in WWI and WWII and discuss the issues they addressed in adapting existing tools to dated and domain-specific language. Another related project with a similar setting is ALCIDE, described in Moretti et al. (2016), a platform that supports the use of text mining techniques for the navigation and visualization of information in historical and literary texts.

3 Dataset

The collection of documents used in this work is composed by 15 printed books, written in Italian, that have been digitized using standard OCR techniques, overall counting over 855,000 words (about 45,000 sentences). The documents are historical memoirs of Italian partisans from the WWII. More specifically, the covered time span goes from the 8th September 1943 to the 25th April 1945, a period known in the Italian historiography as "Resistenza" (Resistance). The geographic area encompassed by the narrated events is the south-western part of the Alps in Piemonte, Italy, with some minor exceptions. The texts have been intentionally selected for digitization for having a partial but significant overlap in terms of narrated events, as well as of places and involved people. None of the 15 documents presents any semantic annotation. Beside the digitization of the documents, three gazetteers have been created: the first one, containing names of persons (1820 entries), has been populated using name indexes provided by 6 of the texts, while the gazetteers containing toponyms and names of organizations (1140 and 190 entries, respectively) have been built manually during the digitization activities. The setting of our work is partly determined by some features of the textual resources under analysis, in particular: 1) due to the specificity of the domain, only 4% of the persons in the gazetteer are available in the italian Wikipedia (according to a manual check carried out on the whole gazetteer); the same problem holds for organizations and, to a smaller extent, for toponyms; 2) while for entities of type Location (LOC) and Organization (ORG) the mining process involves usual problems (abbreviations, upper vs lowercase mention, ambiguity due to the same surface form), with Person (PER) entities the domain at hand presents a further issue as it was quite common, among the partisans, to use aliases, or *nom de guerre*. This feature is showed by 32% of the occurrencies in our PER gazetteer (often the most prominent ones in the narrated events). This means that in text persons are to be found under different combinations of name, surname and nickname. While in some cases this additional information makes the disambiguation process easier, in many other cases it may represent an additional source of ambiguity. The PER gazetteer is structured in three fields, namely Name, Surname and Alias, that are later combined into patterns (see section 4.2); conversely, in the ORG and LOC gazetteers, for each entry all the possible lexical forms are listed (for

	Recognition (%)		
	PER	LOC	ORG
NERD	0.66	0.70	0.51

	Linking (%)		
	PER	LOC	ORG
TagMe	**0.05**	0.45	0.37
NERD	**0.03**	0.47	0.27

Table 1: Evaluation using TagMe and NERD (Percentage of correctly linked occurrencies over a sample of 200 sentences).

the Italian Action Party, for example, we will have: Partito d'Azione, PdA, Pd'A, P.d.A. and so on).

4 Experiment

4.1 Test of existing automatic NERD tools

In order to clarify the need for an ad hoc extraction and disambiguation approach for our texts, we first tried state-of-the-art NERD tools; we randomly selected 200 sentences from the corpus and annotated them with NERD (Rizzo and Troncy, 2012), a framework that aggregates the results from different NER systems (Alchemy API, DBpedia Spotlight, TextRazor, Zemanta among others), and TagMe (Ferragina and Scaiella, 2010), an entity linker to Wikipedia available also for Italian. Table 1 shows the percentage of correctly recognized (i.e. classified) and linked occurrences obtained as result by the two systems. Since TagMe does not separate the two tasks of Recognition and Linking, for this system we only report the Linking results. In the recognition task, NERD performances are quite good for Persons and Locations, while they drop with Organizations. As we turn to the linking task, we observe how the trend in the results is similar in the two systems: performances are very low in the case of Persons, while they improve in the case of Locations and remain quite low for Organizations. This result can partly be explained by the degree of (spatial and social) specificity of the entities that are to be found in the corpus: state-of-the-art tools perform good on prominent entities (for example "Benito Mussolini"), but large-scale knowledge bases lack the suitable knowledge for specific contexts, like those that are more often to be found in the historical memoirs under analysis (and thus NERD systems are not able to link specific entities, such as Chiaffredo Barreri «Tormenta»).

4.2 Methodology

The mining process initially took the form of a simple string matching in text, based on the entries provided by the gazetteers. However, due to the different ways each entity type can appear in text - as discussed in Section 3 - two different strategies have been implemented: string matching with some refinements for LOC and ORG entity types and a slightly more elaborated strategy for PER entities, based on co-occurrence statistics derived directly from the corpus under study.

PER entities. Based on the manual analysis of the documents, 15 lexical patterns have been observed, through which proper names of partisans appear in text; frequent occurring patterns are for example "Name Surname (Alias)", like in "Gustavo Comollo (Pietro)", Name «Alias» Surname, like in "Gustavo «Pietro» Comollo", or "Alias Surname", like in "Pietro Comollo". Each of these 15 patterns have been automatically instantiated for each entry of the gazetteer. This resulted in a dictionary of instantiated patterns that have been used directly for the string matching step in text. Since a certain degree of ambiguity (homonymy) is present in the gazetteer, where many entries share the same name or surname or alias, for each instance of the patterns in the dictionary an ambiguity value has been computed, keeping track, for the ambiguous instances, of all the possible individuals they may actually refer to. For example, the pattern instance "«Renzo»", that in italian can be both a name and an alias, has been connected to all the entries in the gazetteer where "Renzo" appears either as name or as alias, which become candidates for that specific occurrence. Then the string matching in text has been performed. Within the found occurrences, we separated the unambiguous occurrences (those who refer to only one entry in the gazetteer), that have been considered as true positives and did not require further processing, from the ambiguous ones, for which a disambiguation step is needed. Only considering the unambiguous mentions retrieved this way, the system scored a precision measure of .98 (see Section 5), so we used this set of occurrencies as grounding space for the disambiguation step. At this point the system has disambiguated 55.8% (9268) of the PER occurrences in the corpus, while 44.2% (7341) of the occur-

rences remain ambiguous (for precision and recall scores, see Table 2, "Lookup Search"). In order to disambiguate the remaining occurrences different heuristics have been explored. Based on the literature, we tried to apply to the Named Entity Disambiguation task the "one sense per discourse" hypothesis, as done by the authors in (Barrena et al., 2014). Other two heuristics have been explored, that we can informally designate as *Last Mentioned* and *Most Mentioned*. Given an ambiguous occurrence recognized in text, the former one links the occurrence to the last already disambiguated corresponding candidate. Following from the example above, if we find the pattern "«Renzo»" in text, which is ambiguous and corresponds to more candidates from the gazetteer, the system links the mention to the same candidate as the immediately preceding occurrence of this mention. The *Most Mentioned* rule, conversely, assigns to an ambiguous occurrence the candidate which obtained the highest number of mentions in the document. None of these strategies succeeded in improving the performance of the system and this seems to be at least partly due to the length of the documents and to the high ambiguity degree of some entries (consider that the entry "Renzo" alone has 20 candidates in the dictionary, and there are other more ambiguous entries). A promising strategy for the NED task has been individuated using co-occurrence frequencies (Shen et al., 2015; Hachey et al., 2013). Still based on the unambiguous occurrences, for each entry in the PER gazetteer a co-occurrence score has been computed with all the other entities, including Locations and Organizations, at corpus level. The co-occurrence has been considered with other entities in the span of 10 sentences, in terms of raw frequency. Then, given an ambiguous mention and its local context of 10 sentences, the co-occurrence score has been computed for each of its candidates, and the candidate with the highest score has been assigned to the mention. This strategy allows to further disambiguate 10.6% (1764) of the occurrences, with precision and recall scores as indicated in Table 2 ("Lookup Search and Disambiguation").

LOC and ORG entities. For entities of type Location and Organization only the search step has been implemented, not the disambiguation one. However, a cross cleaning has been performed, eliminating nested mentions belonging to different NE categories (for example the name "Leonardo Cocito" in the ORG entity "Battaglione Leonardo Cocito"). In such cases always the longer string has been chosen.

Lookup Search			
	Recall	Precision	F1
PER	0.716	0.980	**0.827**
LOC	0.954	0.917	0.935
ORG	0.987	0.991	0.989

Lookup Search and Disambiguation			
	Recall	Precision	F1
PER	0.751	0.965	**0.845**

Table 2: Evaluation of the presented pipeline.

5 Evaluation

The performances of the system have been evaluated against a manually annotated gold standard made of 1,000 sentences. The gold standard has been built: a) preserving the relative size of each document with respect to the whole corpus size and b) randomly selecting the sentences in a short list that only contains sentences longer than 60 characters and with at least 3 capital letters (which is expected to maximize the probability to have a NE in the sentence). In the resulting gold standard, 1996 entities (belonging to the three mentioned categories) have been annotated as true positives by a single human annotator. The results of the evaluation are presented in Table 2. The co-occurrence approach discussed above allows to gain coverage without losing too much in terms of precision and even if the overall gain is small, the approach shows improvements where other approaches resulted ineffective. The main source of improvement is that, being computed at corpus level, the co-occurrence approach embodies the occurrence information from all the texts, thus going beyond the document level; this proves to be effective when an entity does not appear in unambiguous form in the document at hand but does in other documents of the collection. One limit of the approach emerges when an entity never appears in unambiguous form in the whole corpus, since the grounding space is uniquely based on the set of unambiguous mentions harvested in the search step. Unfortunately this is often the case when memoirs are concerned: many of the authors are non professional writers and do not always provide the

full name of the persons they introduce.

6 Conclusions and Future Works

In this paper we presented an ongoing work aimed at performing Named Entity Disambiguation on a digitized historical corpus, along with the results of the evaluation. Further steps will be a) the refinement of the presented method by means of weighting measures on co-occurrence and possibly of feature optimization techniques, b) the application of the tested disambiguation strategy also to LOC and ORG entities, as well as the study of a cross-category disambiguation strategy, and finally c) the extension of the corpus and of the gazetteers in order to obtain a larger coverage of the domain. Furthermore, this work represents the first step for extracting events and their participants from the presented corpus.

References

Ander Barrena, Eneko Agirre, Bernardo Cabaleiro, Anselmo Penas, and Aitor Soroa. 2014. One entity per discourse and one entity per collocation improve named-entity disambiguation. In *COLING*, pages 2260–2269.

Caroline Barrière. 2016. *Natural Language Understanding in a Semantic Web Context*. Springer.

Federico Boschetti, Andrea Cimino, Felice Dell'Orletta, Gianluca E Lebani, Lucia Passaro, Paolo Picchi, Giulia Venturi, Simonetta Montemagni, and Alessandro Lenci. 2014. Computational analysis of historical documents: An application to italian war bulletins in world war I and II. In *Proceedings of LREC 2014 workshop on Language resources and technologies for processing and linking historical documents and archives - deploying linked open data in cultural heritage (LRT4HDA 2014)*.

Carmen Brando, Francesca Frontini, and Jean-Gabriel Ganascia. 2016. Reden: named entity linking in digital literary editions using linked data sets. *Complex Systems Informatics and Modeling Quarterly*, (7):60–80.

Maud Ehrmann, Giovanni Colavizza, Yannick Rochat, and Frédéric Kaplan. 2016. Diachronic evaluation of ner systems on old newspapers. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016))*, number EPFL-CONF-221391, pages 97–107. Bochumer Linguistische Arbeitsberichte.

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*, pages 1625–1628. ACM.

Anna Goy, Diego Magro, and Marco Rovera. 2015. Ontologies and historical archives: a way to tell new stories. *Applied Ontology*, 10(3-4):331–338.

Ben Hachey, Will Radford, Joel Nothman, Matthew Honnibal, and James R Curran. 2013. Evaluating entity linking with wikipedia. *Artificial intelligence*, 194:130–150.

Giovanni Moretti, Rachele Sprugnoli, Stefano Menini, and Sara Tonelli. 2016. Alcide: Extracting and visualising content from large document collections to support humanities studies. *Knowledge-Based Systems*, 111:100–112.

Federico Nanni, Yang Zhao, Simone Paolo Ponzetto, and Laura Dietz. 2017. Enhancing domain-specific entity linking in DH. *Book of Abstracts of Digital Humanities*, 2:67–88.

Giuseppe Rizzo and Raphaël Troncy. 2012. NERD: a framework for unifying named entity recognition and disambiguation extraction tools. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 73–76. Association for Computational Linguistics.

Kepa Joseba Rodriquez, Mike Bryant, Tobias Blanke, and Magdalena Luszczynska. 2012. Comparison of named entity recognition tools for raw ocr text. In *KONVENS*, pages 410–414.

Wei Shen, Jianyong Wang, and Jiawei Han. 2015. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):443–460.