An ontological approach for locations recognition from Italian non-structured text*

Domenico Cantone, Andrea Fornaia, Marianna Nicolosi-Asmundo, Daniele Francesco Santamaria, Emiliano Tramontana

University of Catania, Dept. of Mathematics and Computer Science
email: {cantone,fornaia,nicolosi,santamaria,tramontana}@dmi.unict.it

Recognizing location names of geographical places and of public or private buildings inside non-structured text documents is an important issue with several practical applications. For example, in the investigative field it is relevant to reveal a place named in the transcription of an interception (i.e., by means of a wiretapping), and in the social media context to reveal, e.g., the places visited by users to provide targeted advertisements.

In the past, this problem has been dealt with with different approaches [11], e.g., using maximum entropy models [8], Conditional Random Fields [5] or automatic learning techniques able to infer the rules for the named entities identification inside a free text [1]. Linked data and ontologies have been used to address the question only in the last decade [7] (a survey of the principal geographical ontologies and datasets can be found in [2]). Many of these approaches, however, are tested, or developed, vertically on top of the English language, making them hard to generalize in order to deal with the Italian language.

In this contribution we focus on the problem of recognizing names of geographical places belonging to the Italian country appearing in non-structured text documents written in Italian. Our approach consists in applying the algorithm introduced in [4] to detect location names inside the text, and semantic web tools such as geographic datasets, *OWL* ontologies [9], and *SWRL* rules [12] to store data and reason on them even in case of name ambiguities. As far as we know, this is the first attempt of handling the problem of location recognition in the context of Italian texts and places using such an approach.

The algorithm presented in [4] is based on a pipe and filter multi-agent model that relies on a set of finite state machines designed from the Italian grammar rules to recognize different sentence patterns where a location name is typically found. It takes as input a non-structured text written in Italian and yields as output a HTML text, where candidate location names have been marked by a label. Then, each location name detected by the algorithm is searched in the *OpenStreetMap* dataset [10] that provides a list of possible matches of real places, each with its related degree of reliability. Finally, the data retrieved by OpenStreetMap are processed to be inserted in a novel ontology, called *OntoLocEstimation*, handling ambiguous geographical names and using the ontology

* Work partially supported by the project *PRIME - Piattaforma di Reasoning Integrata, Multimedia, Esperta* - PON FESR Sicilia 2007/2013 and by the FIR project *COMPACT: Computazione affidabile su testi firmati*, code: D84C46.

OntoLuoghi introduced in [3] that describes in detail the administration of Italian places.

The use of open datasets such as OpenStreetMap allows a widespread and detailed coverage of Italian geographical places, providing a high precision in the detection of real places. This result could not have been achieved using other datasets such as *LinkedGeoData* [6], containing just a proper subset of the data on Italian geographical places stored in OpenStreetMap. In addition, the introduction of SWRL rules permits inferences on knowledge implicitly contained in the novel ontologies which are more refined than the ones allowed in other currently available ontologies on geographical places.

References

1. E. Agichtein and L. Gravano. *Snowball: Extracting relations from large plain-text collections.* In Proceedings of the fifth ACM conference on Digital libraries. ACM, 2000, pp. 85–94.
2. A. Ballatore, D. C. Wilson, and M. Bertolotto. *A Survey of Volunteered Open Geo-Knowledge Bases in the Semantic Web.* Quality Issues in the Management of Web Information, ISRL 50, pp. 93–120, Springer, 2013.
3. D. Cantone, M. Nicolosi-Asmundo, D. F. Santamaria, and F. Trapani. *Ontoceramic: an OWL ontology for ceramics classification.* In Proceedings of the 30th Italian Conference on Computational Logic, CILC 2015, Genova, Italy, July 1-3, 2015, CEUR Workshop Proceedings, ISSN 1613-0073, vol. 1459, pp. 122–127.
4. D. Caruso, R. Giunta, D. Messina, G. Pappalardo, and E. Tramontana. *Rule-based location extraction from italian unstructured text.* In Proceedings of the 16th Workshop "From Objects to Agents", WOA 2015, Naples, Italy, June 17-19, 2015, CEUR Workshop Proceedings, ISSN 1613-0073, vol. 1382, pp. 46–52.
5. J. Lafferty, A. McCallum, and F. C. Pereira, *Conditional random fields: Probabilistic models for segmenting and labeling sequence data.* In Proceedings of the Eighteenth International Conference on Machine Learning (ICML), 2001, pp. 282–289.
6. LinkedGeoData, linkedgeodata.org/.
7. Lyndon J. B. Nixon, R. Volz, F. Ciravegna, and R. Studer. *Ontology based entity disambiguation with natural language patterns.* Fourth International Conference on Digital Information Management, ICDIM 2009, November 1-4, 2009, University of Michigan, Ann Arbor, Michigan, USA, pp. 19–26.
8. K. Nigam, J. Lafferty, and A. McCallum. *Using maximum entropy for text classification.* In IJCAI-99 workshop on machine learning for information filtering, vol. 1, 1999, pp. 61–67.
9. Ontology Web Language, http://www.w3.org/2001/sw/wiki/OWL.
10. OpenStreetMap, www.openstreetmap.org/.
11. S. Sarawagi. *Information extraction.* Foundations and trends in databases, vol. 1, no. 3, pp. 261–377, 2008.
12. Semantic Web Rule Language, http://www.w3.org/Submission/SWRL/.

Authors short biography

Domenico Cantone is professor of Computer Science since 1990, currently at the University of Catania, Italy. He received his Ph.D. degree from New York University in 1987. His main scientific interests include: computable set theory, automated deduction in various mathematical theories, description logic, string matching and algorithmic engineering, and, more recently, rational choice theory from a logical point of view. In the field of computable set theory, he has coauthored three monographs in 1989, 2001, and 2011.

Andrea Fornaia is a PhD student at the Department of Mathematics and Informatics of the University of Catania. In the past has worked on distributed systems technologies and IT infrastructures with the National Institute of Nuclear Physics (INFN) of Catania and the GARR Consortium, the Italian Academic and Research telecommunication network. Current research interests are focused on software quality assurance using software testing, static analysis and code refactoring.

Marianna Nicolosi Asmundo is an assistant professor at the Department of Mathematics and Computer Science of the University of Catania. She received her PhD in Computer Science from the University of Catania in 2003. Her main research interests and activity regard tableau based deduction systems, decision procedures in elementary set theory and non classical logic, description logics and knowledge bases.

Daniele Francesco Santamaria is Ph.D. student at the Department of Mathematics and Computer Science of the University of Catania, Italy. He gradueted magna cum laude in Computer Science with a thesis on computational logic and semantic web. In the past, he was consultant for the Department of Mathematics and Computer Science, working as JAVA software developer and ontologist. His main scientific interests include computational logic, semantic web, algorithmic engineering, software and ontology engineering and developing.

Emiliano Tramontana currently, and since 2007, is an assistant professor of Computing Science at Catania University. His research in the software engineering and distributed systems areas has resulted in more than 100 papers proposing innovative solutions for consistent runtime updating of apps; modular and reusable versions of common design patterns; architectures and algorithms to organise QoS parameters and resources; concepts and metrics for suggesting refactoring opportunities.