

Developing a Semantic Content Analyzer for L'Aquila Social Urban Network

Cataldo Musto^{1,3}, Giovanni Semeraro¹, Pasquale Lops¹, Marco de Gemmis¹, Fedelucio Narducci^{2,3}, Mauro Annunziato⁴, Luciana Bordoni⁴, Claudia Meloni⁴, Franco F. Orsucci⁵, and Giulia Paoloni⁶

¹ Department of Computer Science, University of Bari "A. Moro"

² University of Milano - Bicocca

³ Murex CS s.r.l.

⁴ ENEA - Unita' Tecnica Tecnologie Avanzate per l'Energia e l'Industria - Roma

⁵ Department of Psychology and Language Sciences, University College London

⁶ Department of Humanities and Territory Sciences, University of Chieti-Pescara

Abstract. This paper⁷ presents the preliminary results of a joint research project about Smart Cities. This project is adopting a multi-disciplinary approach that combines artificial intelligence techniques with psychology research to monitor the current state of the city of L'Aquila after the dreadful earthquake of April 2009. This work focuses on the description of a *semantic content analysis* module. This component, integrated into L'Aquila Social Urban Network (SUN), combines Natural Language Processing (NLP) and Artificial Intelligence (AI) to deeply analyze the content produced by citizens on social platforms in order to map social data with social indicators such as *cohesion*, *sense of belonging* and so on. The research carries on the insight that social data can supply a lot of information about latent people feelings, opinion and sentiments. Within the project, this trustworthy snapshot of the city is used by community promoters to proactively propose initiatives aiming at empowering the social capital of the city and recovering the urban structure which has been disrupted after the 'diaspora' of citizens in the so called "new towns".

1 L'Aquila Social Urban Network: a hybrid city model

The city of L'Aquila is located in the mountain region of Abruzzi, in central Italy. In 2009 the life of this town has been seriously modified by an earthquake that caused damage to between 3.000 and 11.000 buildings in the medieval city of L'Aquila. 297 people died in the earthquake. The severe trauma to physical and psycho-social structures is still in the phase of recovery, since the urban centre has not been fully restored and almost half of all citizens still are displaced from their pre-earthquake homes. The specificity of the situation represent a perfect scenario where the principles and the insights behind the concept of

⁷ This paper summarizes the results already presented in the AI*IA 2013 workshop on Smart Cities - URL: <http://ai.unibo.it/SmartCityAIIA2013>

Smart Cities may be applied. Specifically, at ENEA an interdisciplinary team (researchers, architects and engineers) is working on the design of a Social Urban Network (SUN), a hybrid city model [6] developed with the aim of monitoring and revitalizing the traditional social capital and urban heritage with new plans for an integrated future.

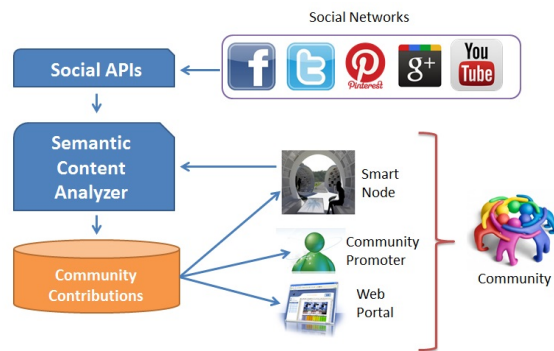


Fig. 1. Social Urban Network (SUN) Architecture

SUN architecture is shown in Figure 1. The input for the whole pipeline is the analysis of the content produced by the citizens on social platforms such as Facebook, Twitter and so on. Next, through a *Semantic Content Analyzer* all the content is deeply processed and analyzed, in order to be map user’s contributions with some well-defined social capital indicators [5] (see Figure 2). Finally, given a *snapshot* of the current social capital obtained through the semantic analysis module, a community promoter can identify activities or specific interventions aimed towards a recovery and empowerment of the social capital of the city through social events, Web portal or a Smart Node, an interactive installation placed in a key area of the city with the aim of creating a meeting place for people that want to share (and to get) information about their town.

2 A semantic content analyzer for the SUN

The semantic analysis module is the core of the whole SUN: it takes as input the information coming from social networks and tries to organize the plethora of data produced by citizens to provide the *community promoters* with valuable information about the current state of the town. The general architecture of the module is shown in Figure 3.

First, a *Social Extractor* exploits Social APIs, such as Facebook⁸ and Twitter’s⁹ ones, to feed a database of community contributions. This database is feed

⁸ <http://developed.facebook.com>

⁹ <http://dev.twitter.com>

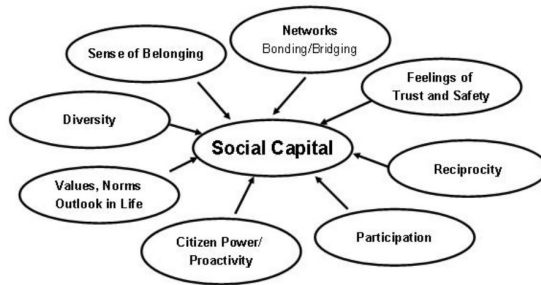


Fig. 2. Social Capital Indicators

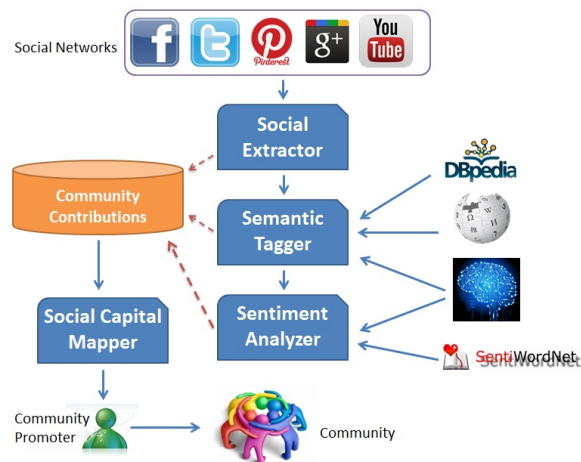


Fig. 3. The architecture of the Semantic Content Analysis Module

according to specific heuristics (e.g. all the tweets containing specific hashtags or coming from a specific geo-location, all the posts crawled from specific Facebook pages, and so on). Next, the database of contributions is processed through three enrichment steps (highlighted with a red dashed arrow): in the first one, a *Semantic Tagger* tries to associate to each piece of content the topic it is about. For this step we will implement an hybrid approach that combines techniques such as LDA with approaches exploiting open knowledge sources (Wikipedia, DBpedia, etc.) such as Tag.me [3] or DBpedia Spotlight [4]. The goal of these techniques is to extract from the text some high-level concept the content is about. As an example, we can consider the tweet in Figure 4. In this case we will process the content of the tweet and we will extract high-level meaningful concepts such as *earthquake* (terremoto, in Italian), *suicides* (suicidi) and *researchers* (ricercatori). Moreover, it is also possible to further connect the concepts with the Wikipedia categories: in this case, for example, 'earthquake' can be connected with the Wikipedia category 'natural disaster'. In this way it is possible to build abstract and high-level connections between different pieces of content written by the community.



Fig. 4. A tweet about L'Aquila

Next, these semantically-enriched pieces of text are processed through *senti-ment analysis* techniques. For this step we will combine machine learning techniques with lexicon-based models that exploit annotated vocabularies (e.g. SentiWordNet [2]) that associate a polarity (positive, negative or neutral) to all the terms of a language. Thanks to these lexicons it is easy to assign a sentiment to the extracted tweets, since it is typically calculated as the weightedsum of the polarity of the terms contained in it. Finally, the Social Capital Mapper builds a classification model able to map each tweet with the social indicators (described in 2) the tweet refers to. Clearly, the social indicator is influenced according to the sentiment conveyed by the tweet. The more positive the sentiment, the higher the social indicator score. Such a simple pipeline, based on the combination of several state-of-the-art machine learning techniques can provide a valuable, meaningful and trustworthy information about people sentiment and opinions.

3 Conclusions and Future Work

In this work we sketched the preliminary design of a semantic content analysis module developed for the SUN of the city of L'Aquila. We figured out a framework where the combined use of techniques for semantic representation and sentiment analysis can help community promoters to rapidly react to people feelings and to design the best initiatives to improve the quality of life for l'Aquila's citizens. However, the project is still ongoing so there is a lot of space for future work: in the next steps most of the effort will be focused on the mapping between the content produced by citizens and the social indicators defined by the psychologists, in order to provide the community promoters with the best possible snapshot of the current situation of the city. Furthermore, we will also work on the comparison of different (semantic) content representation, in order to identify the one able to better represent and convey user sentiments and opinions. Finally, we will integrate a semantic indexer module that implements a word sense disambiguation algorithm [1] and compare the performance of the canonical keyword-based indexing to a more sophisticated semantic one which addresses typical problems related to natural language processing, such as synonymy, polysemy and multi-word expressions.

Acknowledgments. This work fulfils the research objectives of the project PON 01 00850 ASK-Health (Advanced System for the interpretation and sharing of knowledge in health care) funded by the Italian Ministry of University and Research (MIUR)

References

1. P. Basile, M. de Gemmis, A.L. Gentile, P. Lops, and G. Semeraro. UNIBA: JIGSAW algorithm for Word Sense Disambiguation. In *Proc. of the 4th ACL 2007 Int. Workshop on Semantic Evaluations (SemEval-2007)*, Prague, Czech Republic, pages 398–401. Association for Computational Linguistics, 2007.
2. Andrea Esuli and Fabrizio Sebastiani. SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, volume 6, pages 417–422, 2006.
3. Paolo Ferragina and Ugo Scaiella. 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, 2010.
4. Pablo N Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. DBpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*, pages 1–8. ACM, 2011.
5. Franco Orsucci, Giulia Paoloni, Mario Fulcheri, Mauro Annunziato, and Claudia Meloni. Smart Communities: social capital and psycho-social factors in Smart Cities. 2012.
6. Norbert Streitz. Smart hybrid cities: progettare ambienti urbani a prova di futuro. Fondazione Ugo Bordoni, 2010.