Intelligenza Artificiale 6 (2012) 117–119
DOI 10.3233/IA-130039
IOS Press

Guest Editorial

Natural language processing in the web era

Roberto Basili[a,*] and Bernardo Magnini[b]
[a]*Department of Enterprise Engineering, University of Roma, "Tor Vergata", Roma, Italy*
[b]*Fondazione Bruno Kessler, Trento, Italy*

Abstract. Natural Language is still the main carrier for the definition, synthesis and exchange of knowledge in the real world, and this is entirely reflected by Web contents. No interpretation process over Web data is really possible without a more or less explicit reference to natural language(s), the *primordial soup* from which semantics emerges. The advantage and opportunities for NLP research are evident.

This paper introduces the Special Issue of the journal on *NLP in the Web era* by first discussing some opportunities for current NLP research and then summarizing the contribution gathered by the volume.

Keywords: Natural language processing, web applications, corpus-based methods, social web analysis

1. Introduction

Natural Language is still the main carrier for the definition, synthesis and exchange of knowledge in the real world, and this is entirely reflected by Web contents. Although the growing levels of integration, multichannel and modalities of the information made available in the current Web, thus including the Social Web bodies of resources, the central role of language in e-mails, blogs, twits as well as in multimedia pages cannot be denied. Even when multimedia information is made available (as for example, pictures, videos, audio files or digital artworks) natural language is still central as the core vehicle of explanations and complementary crucial information. The role of the annotation processes that enriches these data with linguistic metadata for the localization, retrieval and delivery of the underlying information is evident. No hermeneutic process over such data is really possible without a more or less explicit reference to (possibly multiple) natural language(s), that are thus the "*primordial soup*" from which semantics emerges.

The advantage and opportunities for NLP research are evident. In the Web, sources of rich information *about language* are largely and freely made available. In line with the 90's studies on corpus-driven linguistic knowledge induction and lexical acquisition, several research initiatives (such as, for example the Web as a corpus one [7, 4]) use the Web as the source of useful observation about the lexicon and the syntax so that large scale linguistic knowledge bases can be obtained with reasonable efforts. Moreover, the emergence of novel tasks and applications, such as Classification/filtering, Web search methods, Opinion Mining [8] asks for the adoption of deep language processing methods that are growingly complex. It is also true that the sharing of large scale resources for language processing, from on-line dictionaries to large scale collaborative encyclopedic resources (such as Wikipedia, as discussed in [6]) supports complex forms of induction and linguistic inferences. Finally, it is the Web that promoted large scale benchmarking champaigns (in the spirit of Information Retrieval standard competitions such as TREC), as in the case of the SemEval challenges [3].

*Corresponding author. Roberto Basili, Department of Enterprise Engineering, University of Roma, "Tor Vergata", Via del Politecnico 1, 00133 Roma, Italy. Tel./Fax: +39 06 72597391; E-mail: basili@info.uniroma2.it.

Unfortunately, the increase in volumes also corresponds to a growing complexity in terms of needs, phenomena and applications. On the one side, social media often introduce specific languages that are still largely unexplored by current language processing tools. In fact, the pervasive noise and incompleteness that characterize real documents in blogs, forums or SMS channels also amplify requirements such as coverage and robustness for standard NLP tools. On the other side, while knowledge representation technologies require massive amount of Web data to be traced, linked and semantically harmonized, this whole process is tightly bound by the quality of the linguistic interpretation capabilities that the underlying integration systems can exhibit. Finally, the forms of information retrieval, exchange and sharing used commonly by large communities of Social Web users are such that the semantic management of smaller text units is crucially needed. Specific tasks, such as personalized document management or context-aware search in mobile applications are strongly tight to the interpretation of fine-grain phenomena, such as questions, short queries or twits. In these large scale distributed scenarios multilinguality is also an issue for language processing technologies.

Current challenges are certainly tight to the needs of scaling-up traditional NLP tasks and techniques to Web scale. Examples are the graph based acquisition methods from semi-structured resources, such as Wikipedia [6], the open-domain Information Extraction methods [5] related to the Machine Reading area. On this side the growing interest in distributional models for lexical and compositional semantics corresponds certainly to the perspective of scaling the size and robustness of NLP methods. The task of text similarity recently started at SemEval [2] is an interesting forum for problematic issues, ideas and models of a notion (i.e. similarity) that is central in most Web-based NLP tasks, such as semantic search.

One last point we would like to raise is the role of the Web and its "linguistic" challenges in the scope of what has been called "*a new dawn for AI*" [1]. Renewed interests in AI seems to look at systematic forms of human-like observations, learning and inference in uncertain conditions and dynamically changing environments. Methods such as bayesian statistics and logic are crucially interacting in this perspective, and novel paradigms have been proposed (such as probabilistic programming as in [9]). Notice how the Web represents here a comprehensive environment where software agents can concretely proceed to observations of large scale sets of facts, entities or relations and trigger concrete forms of induction and probabilistic inferences, truly close to a contemporary notion of "*intelligent behaviour*".

As such observations stem from "linguistic artifacts", as those largely populating the Web, the role of natural language is twofold. On the one side, natural language is the privileged medium where the observation occurs: texts in the Web in fact narrate facts about entities and are thus the essence of observable, i.e. existing, things. Every truly AI paradigm along this line has natural language processing (i.e. understanding) as a *core stage*, i.e. the central competence for any AI surviving/existing in the Web world.

On the other side, natural language poses a number of challenges (such as linguistic and world knowledge acquisition or the automation of complex inferences against semi-structured data) that are a natural target for the probabilistic programming AI paradigm. Most results in this latter area will have strong reflections in the AI research literature in the near future. Linguistic challenges are here Semantic Web tasks (in dealing with the Language processing of open linked and ontological data), Web Search processes such as in Question Answering and Intelligent Web Search as well as Language Learning and Knowledge Acquisition. In this perspective, language processing is also relevant for Web Interfaces as well as in the analysis and predictions in Social media applications.

In this special issue five papers have been selected as representative of significant topics about Natural Language Processing in the Web. Among those topics NLP for information access on linked open data, the automatic recognition of entities and thier semantic coreference in the perspective of the Semantic Web, the use of NLP for accessing and combining available web services as well as the extraction of content related to events mentioned in large document archives must be mentioned.

The first paper, by Cabrio and colleagues, addresses the use of NLP for information access in the context of the Web of Data and specifically the Linked Open Data. The paper describes a question answering system over structured data (i.e. DBPedia), exploiting the interpretation of a natural language question based on relational patterns automatically extracted from a corpus. This work shows that both NLP techniques and resources are actually crucial to improve the capability of matching the user need with the continuously growing amount of structured data available on the Web.

The second paper, by Zanoli and colleagues, provides an experimental setting for cross-document

co-reference. As the authors point out, co-reference among entities (e.g. persons, organizations), which are mentioned in different documents is one fundamental aspect of the Semantic Web and of Ontology Population. The paper presents a large-scale experiment on Italian, with the goal of testing different methods for estimating the number of different persons having the same name in a large corpus of news document.

Silvia Quarteroni's contribution on Natural Language Querying of Heterogeneous Data Services addresses the use of NLP for Natural language interfaces to available Web data services. The paper describes a Service Description Framework where both the access to web services and their composition take advantage of a query interpretation process that converts a natural language question into a logical query, through its semantic interpretation.

The paper by Danilo Croce and colleagues on structured learning for Semantic Role Labeling focuses on the relevance of the syntactic structure of a sentence and particularly on the semantic roles as defined by semantic frames. The issue addressed is related to the amount of information that typically machine learning systems need in order to be trained, showing that interesting results can be obtained with a combination of shallow grammatical features and distributional models of lexical semantics. Also in this case the experimental setting is based on datasets developed for the Italian language, showing the portability of the proposed approach.

Finally, Nguyen and Moschitti, on structural reranking models for Named Entity Recognition, focus on improving the state-of-art in Named Entities recognition, a task very much related to semantic-based information access of web resources, such as in the Zanoli et al. paper. More specifically, the paper shows that the introduction of re-ranking techniques based on both flat and structured features improves the performance on Named Entities recognition, for both Italian and English.

Acknowledgments

In the preparation of this special issue the authors are in debt with the support received for the organization of the AI*IA "*Learning by Reading in the Real World*" (LERREW) workshop, held in Palermo on September 15, 2011. Fabio Massimo Zanzotto and Sara Tonelli, as co-organizers of the event, have played a precious role in the success of that event and this special issue.

References

[1] I.A. Ananthaswamy, Algorithm: A new dawn for artificial intelligence, *New Scientist*, 2011.

[2] E. Agirre, D. Cer, M. Diab and A. Gonzalez-Agirre, Semeval-2012 task 6: A pilot on semantic textual similarity, *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012), in conjunction with the First Joint Conference on Lexical and Computational Semantics (* SEM 2012)*, 2012.

[3] E. Agirre, L. Marquez and R. Wicentowski, editors, *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. Association for Computational Linguistics, Prague, Czech Republic, 2007.

[4] M. Baroni, S. Bernardini, A. Ferraresi and E. Zanchetta, The wacky wide web: A collection of very large linguistically processed web-crawled corpora, *Language Re-sources and Evaluation* **43**(3) (2009), 209–226.

[5] O. Etzioni, A. Fader, J. Christensen, S. Soderland and Mausam, Open information extraction: The second generation, In T. Walsh, editor, *IJCAI*, IJ-CAI/AAAI, 2011, pp. 3–10.

[6] E. Hovy, R. Navigli, and S.P. Ponzetto, Collaboratively built semi-structured content and artificial intelligence: The story so far, *Artificial Intelligence* **194** (2013), 2–27.

[7] S.S.A. Kilgariff, editor, *Proceedings of the 7th "Web as Corpus" Workshop (WAC-7)*, 2011.

[8] B. Pang and L. Lee, Opinion mining and sentiment analysis, *Foundations and Trends in Information Retrieval* **2**(1-2) (2008), 1–135.

[9] D.M. Roy, *Computability, inference and modeling in probabilistic programming*, PhD thesis, Massachusetts Institute of Technology, 2011.