An Overview of the Applications of Natural Language to Information Systems

Patricio Martinez-Barco[a,*], Elisabeth Métais[b], Fernando Llopis[a], Paloma Moreda[a]

[a]*Department of Software and Computing Systems, University of Alicante, Apdo. de correos, 99, E-03080 Alicante, Spain*
[b]*Laboratoire Cedric, CNAM, 292 rue Saint Martin, 75141 Paris cedex 3, France*

Abstract

This introduction provides an overview of the state-of-the-art technology in Applications of Natural Language to Information Systems. Specifically, we analyze the need for such technologies to successfully address the new challenges of modern information systems, in which the exploitation of the Web as a main data source on business systems becomes a key requirement. It will also discuss the reasons why Human Language Technologies themselves have shift their focus onto new areas of interest very directly linked to the development of technology for the treatment and understanding of Web 2.0. These new technologies are expected to be future interfaces for the new information systems to come. Moreover, we will review current topics of interest to this research community, and will present the selection of manuscripts that have been chosen by the program committee of the NLDB 2011 conference as representative cornerstone research works, especially highlighting their contribution to the advancement of such technologies.

Keywords: Human Language Technologies, NLP Applications, Information Systems

1. Introduction

In 1995, the International Conference on Applications of Natural Language to Information Systems (NLDB) started a series of meetings aimed at bringing together researchers, industrials and potential users interested in various applications of Natural Language in the Database and Information System area. Day by day, the influence of Human Language Technologies (HLT) on the creation and management of information systems is shown to be more and more obvious. Text mining techniques converge on data mining systems, as well as opinion mining approaches have become an essential part of business intelligence processes. Nowadays, HLT research focuses on Information System applications covering a wide range of areas where NLDB community has great interest in exploring. The exploitation of the Web as a main source on business data systems requires a solid foundation beyond intelligent systems able to automatically retrieve documents where relevant information arises. This relevant information can range from press clipping, where a breakthrough product has been presented by a competency company to social networks, where a community of users, or potential customers, are pouring their opinions about a brand or commercial product. All this information must be filtered, classified, clustered and extracted through Natural Language Processing (NLP) artifacts to finally serve as input to potential business databases. But the task of extracting useful information from unstructured data still poses important challenges. This is the reason why, traditional NLP research areas, such as semantic information retrieval, information extraction, text classification, text mining or question answering systems, achieve prominence with respect to emerging technologies such as opinion mining, subjectivity and sentiment analysis, that are oriented to discover all the knowledge, both objective and subjective, underlying in the Web 2.0. These technologies are destined for becoming the new interfaces to connect future users with modern information systems, where the knowledge will be automatically generated from user interventions in their social networks. In this research framework, the sixteenth NLDB conference was held in Alicante, Spain, bringing together specialized researchers in the field of NLP and information system engineering. In this edition, 74 papers where submitted, and each paper was

[*]Corresponding author. Tel: +34 965909336; fax: +34 965909326

Email addresses: patricio@dlsi.ua.es (Patricio Martinez-Barco), Elisabeth.metais@cnam.fr (Elisabeth Métais), llopis@dlsi.ua.es (Fernando Llopis), moreda@dlsi.ua.es (Paloma Moreda) reviewed by three members of the recommended Scientific Committee. As a result of the review process, 11 articles were accepted as regular papers, 11 as short papers, 23 contributions were selected for the Poster session, and finally, 6 PhD students were selected to present their on-going work in the Doctoral Symposium session held at the conference.

2. Topics of interest for researchers in Natural Language and Information Systems

Despite the different applications and endpoints of each of them, the 74 research papers submitted to the NLDB conference fit with any of the following topics, which were considered as the priority interests of this community:

- Natural Language for Web Information-Intensive Services, that are focused on NLP approaches using semantic and Web technologies oriented to index, classify and extract Web information trying to shift the Internet towards a global platform for the retrieval, combination and utilization of interoperable resources [7].
- Natural Language in Conceptual Modeling, focused on modeling techniques based on NLP to support the adquisition of application domain knowledge through documents and texts analysis [6].
- Natural Language Interfaces for DataBase Querying/Retrieval, trying to build ideal query interfaces to DBs more suitable for occasional users, where there would be no need for the user to spend time learning the systems communication language [1].
- Natural-Language-Based Integration of Systems, mostly of them focused on Ontology-driven Integration and management, where Semantics is considered to be the best framework to deal with the heterogeneity, massive scale and dynamic nature of resources on the Web [8].
- Broadly speaking, any other application, tool, technique or resource of Natural Language Processing when focused to Information Systems, is also considered as priority target.

The selection of manuscripts that have been included in this special issue clearly meets at least three of the five aforementioned topics. However, many of the techniques presented here could have been applied to any of the five areas, and we consider they are truly representative of the work being carried out at this time.

3. Special issue articles and their contributions

The manuscripts contained in this special issue correspond to the extended versions of the four best papers selected after the conference according to the NLDB Program Committee reviews.

1. The first paper, *EmotiNet. A Knowledge Base for Emotion Detection in Text Built on the Appraisal Theories*, extended version of the one in [2]) by Alexandra Balahur, Jesús Hermida, Andrés Montoyo and Rafael Muñoz, from the University of Alicante, covers an important research insight on emotions. Emotions can be reflected by gestures, voice tone, and other forms commonly used by humans to indicate our position or the way in which a concrete event has affected us. However, when communication between two humans is through a text, such as a note or e-mail, there may be a hardly conducive place for the expression of emotions. It is quite possible that emotions of the person issuing the message are different at the time when she/he writes it than at to the time when the receiver reads it. Nevertheless, the main goal of the work performed by Balahur et al., is focused on discovering emotions on the sender of the message when it was issued. Searching emotions in texts is one of the priority challenges where researchers on natural language processing area are pouring all their efforts. This is not a trivial task, since although there exists techniques or procedures in which the issuer can show emotions, such as exclamation marks or emoticons that can already be used in all messaging systems, in practice they are not always used. The appearance in a text of certain words closely associated with emotional states greatly facilitates determining the emotion, ”I'm sad or I'm happy,” is the basic model used by many systems. Emotinet, the proposal of these researchers from the University of Alicante, instead of using traditional models of lexical analysis is based on an analysis of situations and the psychological model of the appraisal theory. The development of this resource has allowed the authors to make a comprehensive evaluation of the proposal as well as a thorough comparison with other traditional systems. Therefore, the main contribution of this paper

resides in showing a novel approach to emotion detection, as well as the interesting conclusions drawn from the comparison with other methods. The good organization and clarity of the paper allows the easy understanding of the proposal, even for beginners in this interesting subject, especially nowadays which written conversations between people who are far away are significantly increasing. This is, definitely, an important contribution to the field of Natural Language for Web Information-Intensive Services

2. The paper *Querying Linked Data Using Semantic Relatedness: A Vocabulary Independent Approach*, extended version of the one in [3]) by Freitas et al., introduces a new method for querying Linked Data on the Web based on the Semantic Relatedness Spreading Activation algorithm using queries in natural language. The method designed for the researches of National University of Ireland (Andre freitas, Joao Oliveira, Sean ORain and Edward Curry) and Universidade Federal de Rio de Janeiro (Joao Carlo Pereira da Silva), introduces a new technique for creating queries using natural language without knowing the underlying structure of the ontologies. To deal with it, authors focused their interests on providing their query mechanism with three basic requirements: usability, flexibility, and expresivity, in addition to the always sought ability to query distributed data. Usability is an intrinsic feature of the natural language queries, therefore, is implicit in the proposal. Moreover, this approach identifies key entities from natural language queries that are mapped to instances or classes in the Linked Data Web by means of a vocabulary-independent matching process, thus allowing the flexibility and expresivity which aimed. Finally, by means of user interaction who filters unrelated results, the mechanism produces valuable outcomes. This mechanism has been implemented and tested on DBpedia proving its value and interest when it is applied to large distributed databases. The exposition of this novel approach, along with the study of related work, makes this paper an important contribution to further research in the field of Natural Language Interfaces for DataBase Querying/Retrieval.
3. The third paper, *Extracting Explicit and Implicit Causal Relations from Sparse, Domain-Specific Texts*, extended version of the one in [4]) by Ashwin Ittoo and Gosse Bouma, from the University of Groningen, deals with the task of mining casual relations from domain specific corpora, in order to discover new knowledge from non-structured databases. Causality is a complex phenomenon through which there is a connection between an agent event (cause of something) and a resultant effect (the principle of cause-effect). The study of cause and effect has always been of great importance in all scientific disciplines. Aristotle (300s BC) introduced the philosophical theory of causality as a way of understanding the human experience of physical nature. Disciplines such as physics, chemistry, medicine and economics, to give some examples, have always based their progress in the discovery of new cause-effect relationships. For this reason, it is worthy to have tools able to extract this knowledge automatically. In this case, authors proposes a minimally supervised algorithm which first adquires a set of reliable causal patterns from a reference broad-coverage corpus (this time, Wikipedia). This is an original way to build a knowledge base, which compensates for the lack of domain-specific resources, and contrasts with other existing techniques because it applies a principled seed selection strategy. This original bet on the use of broad-spectrum corpus to solve problems in restricted domains, the minimized supervision effort, and finally the valuable outcomes obtained after the performance evaluation, outperforming the state-of-the-art, constitute the main contribution of this piece of work on the advancements of Natural Language for Web Information-Intensive Services.
4. Finally, the paper *COMPENDIUM: A Text Summarization System for Generating Abstracts of Research Papers*, extended version of the one in [5]) by Elena Lloret, M. Teresa Romá-Ferri and Manuel Palomar, from the University of Alicante, belongs to the general topic previously cited about Natural Language Processing applications to Information Systems. Specifically, it is an application of a text summarization tool (called COMPENDIUM) to the task of generating abstracts from biomedical papers. Text summarization applications are really quite extensive. While the most obvious is to provide the reader with an overview of a document, it is increasingly being used to feed automatic systems on indexing, search, retrieval, classification and information extraction, because summaries enables them to access the most relevant part of a document without processing it in its entirety. Precisely, because of this secondary effect is why text summarization systems are considered a highly relevant support to information systems. In this case, the authors propose a new method relying on producing an extract that is represented as a word graph, where the shortest paths between nodes are proposed as candidates for new sentences of the abstract. Moreover, the paper presents two different techniques to produce summaries: extractive and abstractive, the former being purely information extraction, the latter including an information compression and fusion stage as a novel approach. The most interesting issue about

this research work is that it does not only focuses on evaluating whether COMPENDIUM produces or not good summaries when applied to biomedical texts, but also analyzes what summarizing technique is more suitable, and the opinion of real users on the assistance provided by this kind of systems to their daily job. To address this latter task, the authors propose a user satisfaction evaluation where users rated the summaries according to their topics, content, and suitability for substituting the original abstracts in the research articles. To sum up, the evaluation results and insights obtained, the novel approach applied to the abstractive technique, and this interesting proposal for evaluation, make this paper a major contribution to the state-of-the-art.

Acknowledgments

First of all, we would like to thank the authors of the contributions submitted to NLDB 2011, and especially to those whose papers where selected for this special issue, for their participation and contribution to the development of research in Natural Language to Information Systems. In addition, we would like to thank the NLDB 2011 Organizing Committee and Program Committee members, as well as their external reviewers, by the contribution to the development of the conference that gave rise to this special issue, and especially to Rafael Muñoz and Andrés Montoyo as Conference Chair and Program Chair, respectively. Finally, we would like to express our gratitude for the financial support given by the Department of Software and Computer Systems at the University of Alicante, the Spanish Ministry of Economy and Competitivity (Spanish Government) by the project grants TEXT-MESS 2.0 (TIN2009-13391-C04-01), LEGOLANG (TIN2012-31224), ATTOS (TIN2012-38536-C03-03), and the Valencian Government (grant no. PROMETEO/2009/119).

References

[1] I. Androutsopoulos, G. D. Ritchie, and P. Thanisch. Natural language interfaces to databases - an introduction. *CoRR*, cmp-lg/9503016, 1995.

[2] A. Balahur, J. M. Hermida, A. Montoyo, and R. Muñoz. Emotinet: a knowledge base for emotion detection in text built on the appraisal theories. In *Proceedings of the 16th international conference on Natural language processing and information systems*, NLDB'11, pages 27–39, Berlin, Heidelberg, 2011. Springer-Verlag.

[3] A. Freitas, J. a. G. Oliveira, S. O'Riain, E. Curry, and J. a. C. P. Da Silva. Querying linked data using semantic relatedness: a vocabulary independent approach. In *Proceedings of the 16th international conference on Natural language processing and information systems*, NLDB'11, pages 40–51, Berlin, Heidelberg, 2011. Springer-Verlag.

[4] A. Ittoo and G. Bouma. Extracting explicit and implicit causal relations from sparse, domain-specific texts. In *Proceedings of the 16th international conference on Natural language processing and information systems*, NLDB'11, pages 52–63, Berlin, Heidelberg, 2011. Springer-Verlag.

[5] E. Lloret, M. T. Romá-Ferri, and M. Palomar. Compendium: a text summarization system for generating abstracts of research papers. In *Proceedings of the 16th international conference on Natural language processing and information systems*, NLDB'11, pages 3–14, Berlin, Heidelberg, 2011. Springer-Verlag.

[6] C. Rolland and C. Proix. A natural language approach for requirements engineering. In *CAiSE*, pages 257–277, 1992.

[7] C. Schroth. The internet of services: Global industrialization of information intensive services. In *Digital Information Management, 2007. ICDIM '07. 2nd International Conference on*, volume 2, pages 635–642, 2007.

[8] A. Sheth and C. Ramakrishnan. Semantic (web) technology in action: Ontology driven information systems for search, integration and analysis. *IEEE Data Engineering Bulletin*, 26:40–48, 2003.

Authors' Bio

Patricio Martinez-Barco obtained his PhD in Computer Science from the University of Alicante (2001). He is working since 1995 in the Department of Software and Computing Science (Language Processing and Information Systems research Group - GPLSI) at this University as Associate Professor, becoming Head of this department between 2009 and 2013. His research interests are focused on Computational Linguistics and Natural Language Processing. His last projects are related to Language Generation, Text and Opinion Mining, Information Extraction and Information Retrieval. He was the General Chair of the ESTAL'04 (Alicante), SEPLN'04 (Barcelona), and IBEREVAL'2010 (Valencia), and co-organized serveral workshops and conferences related to these topics. He has advised four PhD Thesis, has edited several books, and contributed with more than 80 papers to several journals and conferences. Currently, he is Vice-President of the Spanish Society for Natural Language Processing (SEPLN).

Elisabeth Métais is a Full Professor at the CNAM University (Paris, France) and a researcher in the CEDRIC Laboratory since September 2000. Up to 2000 she was an Associate Professor at the University of Versailles (France) working in the PRiSM Laboratory and she previously was a researcher for the University of Paris VI (France) where she holds her Ph.D. in Computer Science (1987). Her main axe of research has been Database Design. She participated in the definition of SECSI, the first expert system in database design. She has been interested since the early nineties in applying Natural Language Processing techniques to Database Design and initiated the NLDB (applications of Natural Language to Data Bases and information systems) in Versailles in 1995. She is currently interested in memory prosthesis and memory-aids for elderly persons and Alzheimer disease patients; she is managing a project concerning the help of ICTs and smart cities for Alzheimer patients living in the City of Paris.

Fernando Llopis received his Masters degree in Computer Science at the Polytechnic University of Valencia, later obtained his Ph.D. in the University of Alicante. He is Associate Professor of this University and his main teaching areas focuses on the analysis, design and development of applications. He is working since 1993 in the Department of Software and Computing Science (Language Processing and Information Systems research Group - GPLSI) at this University. His research interests are Human Language Technologies (HLT) and Natural Language Processing (NLP), in particular Information Retrieval and Questions Answering. He was the Dean of the High Polytechnics School of the University of Alicante during 2005-2013 He is the author of over 75 scientific publications in relevant journals and international conferences and several books about programming. .

Paloma Moreda is an Associate Professor of the University of Alicante in the Department of Software and Computing Science in which she is currently the Deputy Head. She obtained her PhD in Computer Science by this

University in 2008. As a member of the Natural Language Processing and Information System research Group (GPLSI) her main research areas of interest are focused on Computational Linguistics and Human Language Technologies, concretelly semantic analysis, informality and text normalization, and simplification of texts. She is currently the IP at University of Alicante for the Flexible Interactive Reading Support Tool european project (FIRST). She has collaborated in the organization of several conferences and workshops related with her research areas and contributed with more than 50 scientific publications in international relevant journales, conferences and whorkshops.