Introduction to the issue: From Natural Language Processing to Information and Human Language Technology

Vera Lúcia Strube de Lima, Carlos Augusto Prolo

Pontifícia Universidade Católica do Rio Grande do Sul
Av. Ipiranga, 6681, prédio 32
Porto Alegre, RS (Brazil)
{vera.strube,carlos.prolo}@pucrs.br

Interest in Natural Language Processing (NLP) basically starts at the origin of Computer Science, with the aspects chosen by Alan Turing in his proposal to define intelligent behaviour, where the capacity to cope with natural language is stressed. A key issue in boosting the area was the military applications, during the cold war, leading to emphasis in projects of automatic translation and its applications. At the time limited success was achieved given the expectations, but they were important to a repositioning, followed by a valorization of more fundamental issues, such as knowledge representation, and subsequently a relative equilibrium between basic research and applications.

Natural language processing arrives to the 21st century with new and renewed concerns [3]. The concerns with the evaluation of the developed systems and with large-scale language processing are some of them. The insertion of the practical results of this area in daily systems has become a must.

The demand for more concrete results is strongly connected to: the possibility of access in natural language to large knowledge bases over multiple domains, with suitable interfaces and intelligent search engines; the representation of concepts and vocabulary and the interoperability of the structures used for this representation; the automatic document indexation, summarization and categorization; automatic translation itself; and the several challenges arising from access universalization, multilinguism and multiculturalism [1].

In order to develop such applications, however, theoretical and practical studies still need to be developed, and more than that, a union of efforts among several research fields is required, even if there are broader challenges to be answered by cognitive sciences, linguistics and neurosciences. In this context, the research in natural language processing is inserted under the denomination of information and human language technology.

Information and human language technology has crucial relevance in the current stage of development of human society. Natural language processing made use of a rationalist approach [2] from the 1960´s to the 80´s, characterized by the belief that a "significant part of the knowledge in the human mind is not derived by the senses, but is fixed in advance, presumably by genetic inheritance." In Artificial Intelligence, this belief led to the construction of systems with a good deal of *hard encoded* knowledge, and associated reasoning mechanisms. The empiricism that came right after attenuated this approach: "no learning is possible from a completely blank slate, a *tabula rasa* … Rather, it is assumed that a baby's brain begins with general operations for association, pattern

Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial. Vol 11, No.36 (2007), pp. 5-7.
ISSN: 1137-3601. © AEPIA (http://www.aepia.org/revista).

recognition, and generalization, and that these can be applied to the rich sensory input available to the child to learn the detailed structure of natural language." This approach, based on the analysis of large volumes of samples organized as corpora, is leading to the application of statistical methods, probabilistic models, pattern recognition, and machine learning in general.

Information and human language technology is an inherently multidisciplinary area. As such, its development requires a coordinated effort of integration of diverse communities with different skills, knowledge and research methods. The Workshop on Information and Human Language Technology – TIL (the acronym stands for *Tecnologia da Informação e Linguagem Humana*), – was conceived in Brazil to provide a forum to integrate these communities, exchanging knowledge, research and methods, creating an environment of collaborative work towards the common objective of developing information and human language technology. Its first edition was held in 2003, in São Carlos, state of São Paulo, and since then TIL has become an annual event.

Conceived originally to boost research in Portuguese in particular, in its fourth edition TIL 2006 was held in Ribeirão Preto, São Paulo, in October 26-27, collocated with IBERAMIA, the Ibero-American Artificial Intelligence Conference. This provided the opportunity to extend its reach to related languages, in particular Spanish. Researchers, faculty members, students, and professionals, from diverse disciplines were brought together, to present their contributions from Computer Science, Linguistics, Information Science, Electrical Engineering and Psycholinguistics, among others. Contributions reporting concluded or ongoing work were submitted in the form of regular papers, and the accepted papers were published in the Proceedings of the Workshops of the IBERAMIA-SBIA-SBRN 2006, on CD-ROM.

TIL 2006 topics of interest included: Natural Language Processing (Parsing, Part-of-Speech Tagging and Morphological Analysis, Spoken Language, Language Generation, Semantic Analysis, Discourse Models, and Co-reference); Linguistics (Phonetics and Phonology, Syntax and Morphology, Lexicology and Lexicography, Semantics, Pragmatics and Discourse, Corpus Linguistics, Psycholinguistics, Grammar Design and Inference, and Linguistic Theories applied to NLP); Statistical and Corpus-Based NLP; and NLP Applications (Machine Translation, Information Retrieval, Classification and Clustering, Data Mining, Language Ontologies, Question Answering, Summarization, Language-based Human-Computer Interfaces, and Practical Systems Construction and Evaluation).

There were 53 paper submissions: 22 of them were accepted for oral presentation and publication in the proceedings. Concerning the first author, 13 of the accepted papers had a Brazilian first author, whereas in the remaining 9, the first author was from the following countries: Argentina, Denmark, Finland, France, Mexico, Portugal and Spain. Among the accepted papers, 17 were effectively presented during the Workshop, to 68 registered attendees.

This special issue of the *Revista Iberoamericana de Inteligencia Artificial*, contains the eight best papers of TIL 2006, selected according to the evaluation score from the original review process, but also considering the impact and discussions during the presentations at the event. The revamped versions have undergone a separate revision for the issue.

The first two papers are on more conceptual issues. On the computational side, *The obligations and common ground structure of practical dialogues* is related to Discourse and presents a theory of dialogue structure for task oriented conversations. On the linguistic side, *Adjectival Semantics and the Legal Domain: a study for ontology improvement* is a study of the semantic of the adjectives used in the legal domain towards the construction of an ontology of concepts which is argued to be useful for Information Retrieval systems.

The remaining papers are more directly oriented to application systems. In *Designing topic shifts with graphs* the application the authors have in mind is to help users with certain initial background knowledge, given in the form of a topic, to move forward to acquire knowledge in another topic. Given such pair of topics, their goal is to return what would be a suitable reading list that would allow for the user to move smoothly from one topic to the other. Search space is modelled as a graph of documents with topic distances.

Lexicon Construction for Information Systems is an NLP application to software engineering. The authors aim at building a shared lexicon of the application domain from the documents pertaining

to the requirements engineering phase of software development.

Extract-biased pseudo-relevance feedback presents an approach to refine Information Retrieval (IR) queries using the technique of pseudo-relevance feedback, based on extracts of documents instead of the whole documents.

There are two articles focusing on document authoring. *Some issues on complex networks for author characterization* reports experiments concerning the applicability of complex networks to author characterization, investigating the correlation between several properties of the networks used to model texts of different genres and authors, with corresponding authors' characteristics. *Author identification using stylometric features* presents a writer-independent method for author identification, and a stylometric feature set to be used with the algorithm which is argued to be suitable for Portuguese texts.

Finally, focusing on a more specific task, frequently embedded in several application systems, *Machine Learning algorithms for Portuguese Named Entity Recognition* proposes several Named Entity Recognition (NER) machine learning algorithms, applying them to Portuguese and evaluating their performance.

References

[1] D.S Jurafsky and J.H. Martin. Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition. Prentice-Hall, Uper Saddle River, NJ, US. 2000.

[2] C Manning and H. Schütze. Foundations of statistical natural language processing. MIT Press, Cambridge, MA, US. 1999.

[3] V. L. Strube de Lima, M. d. G. V. Nunes, and R. Vieira. Desafios do processamento de línguas naturais. In: Proceedings of the 34º Seminário Integrado de Software e Hardware (SEMISH 2007). Rio de Janeiro, Brazil. 2007.