Introduction to the 3rd ROMAND workshop on Robust Methods in Analysis of Natural Language Data

Vincenzo Pallotta
Swiss Federal Institute of Technology – Lausanne, CH
University of California – Berkeley, CA
Vincenzo.Pallotta@epfl.ch

Amalia Todirascu
University of Troyes - France
University "I.A. Cuza" of Iasi - Romania
Amalia.Todirascu@utt.fr

Robustness in Computational Linguistics has been recently recognized as a central issue for the design of interactive applications based on natural language communication. If a failure of the system can be acceptable in batch applications requiring a human intervention, an on-line system should be capable of dealing with unforeseen situations in a more flexible way. When we talk about system failure we do not think at inherent program failures like infinite loops or system exception, we consider, rather, failures related to the processing of the input and its assimilation in the system's knowledge base. A failure of this kind means simply that the system does not "understand" the input. The automated analysis of natural language data has become a central issue in the design of Intelligent Information Systems. Processing unconstrained natural language data is still considered an AI-hard task. However, various analysis techniques have been proposed in order to address specific aspects of natural language. In particular, recent interest has been focused on providing approximate analysis techniques, assuming that when perfect analysis is not possible, partial results may be still very useful.

Interpretation of natural language data is a subjective cognitive process. Its formalisation could be either a straightforward or a hard task, depending on the perspective taken. The human ability to interpret language is the result of thousands of years of evolution and cultural development. We, as humans, are capable of *mapping* surface language forms into meaning representations, but we can only observe the final result of this process without exactly knowing what happens in our brain. In contrast, we can model understanding in an abstract domain where the process of reaching the meaning is decomposed at least in two parts. We are interested in establishing to what degree:

- the mapping is sound (i.e. if the competence model enables us to extract correct meanings);
- the mapping is complete (i.e. if the competence model enables us to deal with all the language phenomena).

These two measures are fairly general, but in specific applications they might correspond to the classic evaluation metrics of Information Retrieval (i.e. precision and recall).

Even human interpretation of language is not infallible. For instance, in areas where people lack context, or have different views on the context, people can fail to understand each other and can have different opinions on utterances' interpretation. Apparently, it is always possible to provide an interpretation of any kind of data. Actually, one not always provides the right or the best interpretation among the possible ones. This happens for humans and, why not, for artificial systems. When switching to artificial systems, what is worth considering is the human ability to provide different degrees of approximate interpretations, ranging from the full understanding to the complete ignorance. In addition, humans are able to overcome their limitations by their learning capabilities. In the interpretation process, the lack of knowledge, the uncertainty, vagueness, ambiguity, and misconception should be explicitly represented and considered at a meta-level in order to handle linguistic applications 'robustly'.

There are many ways in which the topic of robustness may be tackled: as a competency problem, as a problem of achieving interesting partial results, as a shallow analysis method, etc. What these approaches have in common is that the simple (rigid) combination of 'complete' analysis modules at different linguistic levels does not guarantees the system's robustness. Rather, robustness must be considered as a system-wide concern. We believe that the problem of robustness in NLP may be tackled by adopting the following two complementary approaches:

1. as an engineering `add-on': completing an existing system with additional features in order to overcome the problem of its inability to cope with real-world data;

2. as a basic element of the underlying language theories: extending them by assuming that the understanding of the domain can be incomplete.

Both approaches may be effective under certain circumstances. We thus propose to consider two different perspectives on the role of robustness in software architectures for natural language processing and understanding, namely: robustness "in the small" and "in the large".

With *Robustness "in the small"* we mean the robustness in language analysis is achieved at the individual linguistic levels, such as the morpho-syntactic analysis, semantic interpretation, conversational analysis, dialogue acts recognition, anaphora resolution, and discourse analysis.

With *Robustness "in the large"* we mean the robustness achieved in integrated NLP/NLU architectures possibly implementing hybrid approaches to language analysis, and incorporating the different methods into a competitive/cooperative system.

ROMAND 2004 is the third of a series of workshops aimed at bringing together researchers that work in fields like artificial intelligence, computational linguistics, human-computer interaction, cognitive science and are interested in robust methods in natural language processing and understanding. Theoretical aspects of robustness in Natural Language Processing (NLP) and Understanding (NLU) are concerned by the workshop's theme, as well as engineering and industrial experiences.

This volume contains an extended abstract of the invited talk and 11 papers selected by peer review out of 16 submissions. The accepted papers cover topics related to robust syntactic parsing, robust semantic parsing and applications using robust analysis methods (semantic tagging, information extraction, question answering, document clustering).

The third edition of ROMAND workshop features an exceptional invited speaker. Frank Keller from Edinburgh University accepted to talk about robustness aspects in cognitive, computational and stochastic models of human parsing surveying and discussing the weaknesses and strengths of most recent advanced theories.

The papers dedicated to robust syntactic parsing methods cover topics as combinations of statistical and deep-linguistic syntactic analysis, as well as a parser's evaluation. The paper proposed by G. Schneider, J. Dowdall and F. Rinaldi, "A Robust and Hybrid Deep-Linguistic Theory Applied to Large-Scale Parsing", presents an efficient state-of-the-art hybrid parser combining statistical and rule-based parsing as well as shallow and deep parsing using a combination of phrase-structure and functional dependency grammars. The paper "Syntactic parser combination for improved dependency analysis" describes how F. Brunet-Manquant improves parsing efficiency in building complex dependency structures by combining the results of the three concurrent parsers: the Incremental Finite-State Parser, the GREYC parser combining tagging methods to build non-

recursive chunks, and the Xerox Incremental Parser. A difficult problem of evaluation as well as the evaluation of the GETARUNS system is proposed and discussed in Delmonte's paper "Evaluating GETARUNS Parser with GREVAL Test Suite".

Robustness plays an important role in semantic interpretation. Semantic interpretations are generated by robust syntactic parsers output in specific representation languages: as logical formulae in Minimal Recursion Semantics or as semantic hypergraphs in Unified Network Language respectively in the papers "A step towards incremental generation of logical forms" by L. Coheur, N. Mamede, G. Bés, and "Using an incremental robust parser to automatically generate semantic UNL graphs" by N. Gala. Existing knowledge bases (FrameNet and WordNet) are exploited to build complex semantic structures directly from free texts, as proposed in "An Algorithm for Open Text Semantic Parsing" by L. Shi and R. Mihalcea. J. Bryant in his paper "Recovering Coherent Interpretations Using Semantic Integration of Partial Parses" presents a robust semantic parser for Embodied Construction Grammars used to reconstruct full semantic interpretations from semantic chunks in the framework of psycholinguistics studies on language acquisition.

A direct use of robust analysis methods is featured by several Information Extraction applications: Part of Speech tagging, building knoweldge maps from texts, Question-Answering, and document clustering. Robust morphological analysis for POS and restricted semantic tagging in Bulgarian is achieved by for learning robust ending guessing rules in the P. Nakov and E. Paskaleva's paper "Robust Ending Guessing Rules with Application to Slavonic Languages". Extraction of Associative Term Networks from texts including co-occurences of several content words sharing similar contexts is the goal of the paper "Knowledge Extraction Using Dynamical Updating of Representation" by A. Dragoni, L. Lella, G. Tascini, and W. Giordano. Answer Validation is an important module of a Question-Answering system for which a method exploiting co-occurrence frequencies of keywords extracted from Web documents is proposed in the paper "Answer Validation by Keyword Association" by M. Tonoike, T. Utsuro, and S. Sato. Document clustering algorithms could help the users for Web browsing, where several document representations are compared (as POS or as WordNet synsets) and exploited by an efficient clustering algorithm discussed in the paper "WordNet-based text document clustering" by J. Sedding and D. Kazakov.

We believe that the output of the ROMAND 2004 workshop will contribute to a better understanding of various aspects of robust analysis in Natural Language Processing and Understanding by presenting relevant advances in morphology, syntax, semantics, pragmatics, and evaluation, as well as examples of large-scale Information-Extraction applications relying on robust NLP/NLU techniques and architectures.

We would like to thank all the people who have supported the 3rd edition of ROMAND, in particular the authors who submitted their works, the members of the scientific program committee, the COLING workshop program committee, and the local organizing staff.