Natural Language Engineering **8** (2/3): 93–96. © 2002 Cambridge University Press
DOI: 10.1017/S1351324902002942 Printed in the United Kingdom

Robust methods in analysis of natural language data

AFZAL BALLIM and VINCENZO PALLOTTA
School of Information and Communication Sciences, Swiss Federal Institute of Technology, Lausanne, Switzerland

(*Received 7 May 2001*)

Introduction

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 as an AI-hard task. However, various analysis techniques have been proposed 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 upon the perspective taken. The human ability to interpret language is the result of thousands of years of evolution and cultural development.

Human interpretation of language is not infallible, either, for instance in areas where people lack context or have different views on the context, people can fail to understand or have differences of opinion on interpretation. It is apparent that it is always possible to provide an interpretation of any kind of data. In contrast, one is not always able to provide 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 interpretation, ranging from full understanding to complete ignorance. In addition, humans are able to overcome their limitations by their learning capabilities.

Lack of knowledge, uncertainty, vagueness, ambiguity, and misconception should be represented and considered in the interpretation process to handle linguistic applications 'robustly'. This will enable us to design an open system where missing competences can be plugged-in to improve its problem solving capabilities.

This special issue presents the state-of-the-art in methods for robust natural language processing and understanding. The following computational approaches to natural language analysis are motivated by the need for modelling the systems' behaviour not solely by means of exclusively general principles, but also by means of content information. Moreover, the coexistence of multiple theories may help the

https://doi.org/10.1017/S1351324902002942 Published online by Cambridge University Press

systems in finding the most appropriate analysis strategy (e.g. heuristically choosing which modules to use and how to compose them). Within a variety of tuning parameters that may radically alter their performance and outcomes. The ability to determine a-priori what are the parameters that are responsible for different types of analysis is a key factor for achieving adaptive NLP processors.

Most often, developed systems have been committed to a closed world assumption; a failure in fulfilling a goal or in providing a solution to a given problem has always been interpreted in a negative way: as an impossibility rather than a temporal inability due to the actual lack of specific knowledge. Moreover, the lack of knowledge may be considered itself as meta-knowledge and its awareness may help in judging the quality of the provided solution.

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 they have in common is that no simple combination of 'complete' analysis modules for different linguistic levels in a chain can give a robust system, because they cannot adequately account for real-world data. Rather, robustness must be considered as a system-wide concern.

In summary, 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 a system with additional features to take into account the inability of our theories to cope with real-world data;
2. as a basic element of our theories: developing theories which assume that understanding of the domain can be incomplete.

Both approaches may be valid under certain circumstances.

Objective of this special issue

The main goals of this special issue of the *Natural Language Engineering* journal are devoted to advances in fields like artificial intelligence, computational linguistics, human-computer interaction, and cognitive sciences which are faced with the problem of feasible and reliable NLP systems implementation. We consider of central interest improving and integrating various processing methods with respect to the following issues:

- Extending coverage.
- Improving efficiency.
- Disambiguation ability.
- Approximate processing.
- Enhancement of underlying theories.

After the *ESSLLI'96 Workshop on Robust Parsing*, organised by John Carroll and Ted Briscoe, the only initiatives aimed at gathering contributions about robust methods in NLP were the ROMAND 2000 workshop organised and the collection of papers edited by Jean-Claude Junqua and Gertian van Noord on 'Robustness in Language and Speech Technology'. As in the case of the ROMAND workshop, we

https://doi.org/10.1017/S1351324902002942 Published online by Cambridge University Press

decided to expand our vision of robustness in NLP by considering other outstanding issues beyond syntactic parsing and speech recognition.

In this special issue we have investigated additional aspects of NLP such as semantic, discourse, and NLP architectures other than those considered in the ESSLI'96 workshop and the collection edited by Junqua and van Noord. This special issue has a wider spectrum of interests and is in line with the main motivations of the ROMAND workshop where the participants showed a growing interest in widening the scope of applications for robust methods in NLP.

The articles

The eight articles included in this special issue all address the concerns above, in the context of robust methods in analysis of natural language data. The papers naturally fall into one or more of the following four conceptual groupings:

1. Design methodology and performance evaluation for robust analysis systems;
2. Robust syntactic parsing;
3. Semantic and pragmatic content analysis;
4. Language engineering architectures.

In their paper, Roberto Basili and Fabio Massimo Zanzotto propose a framework for the systematic design and evaluation of robust parsers based on the notion of empirical evaluation of their performance. They provide a useful methodology for the rational design of a robust parsing system, together with a useful metric for performance evaluation which is based on an original quantitative model. As an example of their methodology, the CHAOS system is presented and evaluated.

The contribution of Ait-Mokhtar, Chanod and Roux is focused on how to design a robust deep parser by means of a multi-layered architecture in which each layer safely applies rules to enrich the input incrementally with linguistic descriptions at different linguistic levels. They propose both a methodology to achieve robust deep parsing, and an application of their principle to the deep parsing of French.

The work of Ciravegna and Lavelli follows similar principles, but focuses more on its application in Information Extraction. In particular, full parsing (i.e. which goes up to a semantic form) is obtained by a cascade of finite state-based linguistic processors. One important issue here is that computational overhead is kept minimal since typical performance bottlenecks in a full parser are avoided in their approach by adopting several semantic-driven heuristics to prune the search space and by building a full parse structure relevant for the Information Extraction task. They also propose a related framework for grammar management.

An adaptive Information Extraction system is proposed by Turmo and Rodriguez which exploit the ELVIUS learning module to react robustly to the changes of domain or language in the context of the TURBIO system. The TURBIO system uses rules aimed at mapping linguistic information into multiple concepts of the target 'scenario of extraction'. These rules are induced from an annotated training corpus using a Multi-strategy Constructive Learning approach.

https://doi.org/10.1017/S1351324902002942 Published online by Cambridge University Press

In the article by Takenobu, Kenji, Hironori and Hozumi, machine learning techniques are applied to improve the quality of information retrieval (i.e. improving the specificity of index terms without losing exhaustivity) by the construction of complex index terms using linguistic knowledge and their smart selection using a decision tree. Selected complex terms are used to re-rank conventional retrieval results. The latter are obtained by means of simple index terms.

A robust approach to Topic Analysis is investigated in the article by Olivier Ferret and Brigitte Grau. They propose a 'bootstrapping' approach in which the result of a topic analysis based on automatically extracted knowledge is fed to a module for building topic representations to be in turn exploited for improved further analyses. The knowledge used in the second stage is learned from a segmented input resulting in improved overall precision.

Frank Schilder proposes an algorithm for discourse parsing that employs a formal discourse grammar which is based on underspecification, and techniques from Information Retrieval (IR) applications to derive the discourse structure robustly. The IR techniques compensate for the lack of a world knowledge reasoning system and are used for computing topicality scores helping to predict discourse attachments. The underspecification ensures the independence of the algorithm on any specific discourse theory and its flexibility, such that further developments can encode more specific constraints on the discourse structure.

A multi-purpose Named Entity recognition system, the MUSE system, is presented in the article by Maynard, Tablan, Cunningham et al. They focus their attention on how robustness in Language Engineering systems can be tackled from the point of view of software engineering. The GATE architecture is proposed as a valuable contribution to engineering robustness since it enables the rational design of NLP applications and a methodology for their sound implementation from specifications. The notion of predictability in software construction is highlighted as a major factor for the engineering of robust NLP applications.

Acknowledgements

We wish to thank everybody who supported our initiative and in particular Prof. John Tait for his help and supervision during the revision process. We would like to express a special thanks to the reviewers, whose views and comments were of great aid to the editors as well as the authors, and without whom this issue would not have been possible. We would also like to mention the high quality of the scientific committee and thank all the members for their participation in our initiative. We would also like to thank Professor Giovanni Coray for his encouragement in pursuing our vision on robust methods in NLP.

https://doi.org/10.1017/S1351324902002942 Published online by Cambridge University Press