Lecture Notes in Artificial Intelligence 1714

Subseries of Lecture Notes in Computer Science
Edited by J. G. Carbonell and J. Siekmann

Lecture Notes in Computer Science
Edited by G. Goos, J. Hartmanis and J. van Leeuwen

Springer
Berlin
Heidelberg
New York
Barcelona
Hong Kong
London
Milan
Paris
Singapore
Tokyo

Maria Teresa Pazienza (Ed.)

Information Extraction

Towards Scalable, Adaptable Systems

Springer

Series Editors

Jaime G. Carbonell, Carnegie Mellon University, Pittsburgh, PA, USA
Jörg Siekmann, University of Saarland, Saarbrücken, Germany

Volume Editor

Maria Teresa Pazienza
Department of Computer Science, Systems and Production
University of Roma, Tor Vergata
Via di Tor Vergata, I-00133 Roma, Italy
E-mail: pazienza@info.uniroma2.it

Cataloging-in-Publication data applied for

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Information extraction : towards scalable, adaptable systems / Maria Teresa Pazienza (ed.). - Berlin ; Heidelberg ; New York ; Barcelona ; Hong Kong ; London ; Milan ; Paris ; Singapore ; Tokyo : Springer, 1999
(Lecture notes in computer science ; 1714 : Lecture notes in artificial intelligence)
ISBN 3-540-66625-7

CR Subject Classification (1998): I.2, H.3

ISBN 3-540-66625-7 Springer-Verlag Berlin Heidelberg New York

Typesetting: Camera-ready by author
SPIN: 10705092 06/3142 – 5 4 3 2 1 0 Printed on acid-free paper

Preface

The ever-growing interest in new approaches to information management is strictly related to the explosion of collections of documents made accessible through communication networks. The enormous amount of daily available information imposes the development of IE (Information Extraction) technologies that enable one to:

- access relevant documents only and
- integrate the extracted information into the user's environment.

In fact, the classic application scenario for IE foresees, for example:

1. a company interested in getting detailed synthetic information (related to predefined categories);
2. the documents, as sources of information, located in electronically accessible sites (agencies' news, web pages, companies' textual documentation, international regulations etc.);
3. the extracted information eventually being inserted in private data bases for further processing (e.g. data mining, summary and report generation, forms filling,...).

A key problem for a wider deployment of IE systems is in their flexibility and easy adaptation to new application frameworks. Most of the commonly available IE systems are based on specific domain-dependent methodologies for knowledge extraction (they ignore how to pass to templates related to other domains or different collections of documents). The need exists for more principled techniques for managing templates in a domain-independent fashion by using the general structures of language and logic. A few attempts have been made to derive templates directly from corpora. This process is similar to deriving knowledge structures and lexicons directly from corpora. This methodological approach (*adaptability*) could push for a rapid customization to new domains of existing IE systems.

The missing availability of robust natural language processing (NLP) tools is an obstacle in developing efficient systems for information management and broadcasting.
The use of written texts as sources of knowledge lags behind other applications: it is crucial to characterize the suitable framework to support and simplify the construction phase for NL-based applications. The present software engineering methodologies are not adequate, while the automatic manipulation of unstructured natural language texts will become an important business niche.

Information Extraction technology is required to get performance levels similar to Information Retrieval (IR) systems proved to be commercially viable.

In many respects, IR and IE are very often used with a similar meaning when the interest is in *extracting*, from a very large collection of textual documents, useful information matching linguistic properties. Likewise, the Message Understanding Conferences (MUC) and the Text Retrieval Conferences (TREC) are the most qualified environments in which different IE and IR approaches, respectively, are evaluated with respect to the ability of identifying relevant information from texts. In both these competitions, innovative approaches have been implemented, evidencing the role of NLP systems. Nevertheless the definition of how accurate an approximation to explicit linguistic processing is required for good retrieval performances is still under debate.

Multilingual information extraction (IE) methodologies are more and more necessary.
Even if the most common language used in electronic texts is English, the number of languages adopted to write documents circulating and accessible through networks is increasing. Systems developed for such an application must rely on linguistic resources being available in several languages. Traditionally, these resources (mainly lexicons) have been hand-built at a high cost and present obvious problems for size extension and portability to new domains.

Most of the resources needed for IE systems are still developed by hand.
This is an highly time consuming task for very expensive human experts. A possible solution is in extracting linguistic knowledge from corpora. This requires developing systems that, in a unified approach, would be able to

1. extract such linguistic knowledge, and
2. represent it, preferably at a meta-level independently from source language and application domain.

Parallel corpora may be considered as valuable sources of this meta-knowledge, in case aligned multilingual parallel corpora are available and tools for equivalent processing have been developed.
Alignment in these texts is mandatory and it must be verified at some level (at least paragraphs and sentences). Two different frameworks exist for this task:
- use of some sort of traditional linguistic analysis of the texts, or
- a statistical approach.
The former seems to be based on the same kind of information they are trying to extract. The latter, based on simpler assumptions (e.g. a significant correlation exists in the relative length of sentences which are translations of each other), is currently used.

All these themes will be analyzed and debated at SCIE99, the SChool on Information Extraction, organized by the Artificial Intelligence Research Group of the University of Roma Tor Vergata (Italy) and supported by the European Space Agency (ESA), the Italian Association for Artificial Intelligence (AI*IA) and the National Institution for Alternative Forms of Energy (ENEA).

In recent years, SCIE99 (the second conference, SCIE97 being the first) appears to to have become an important forum in which to analyze and discuss major IE concerns.

By comparing the lectures held at the School on Information Extraction, SCIE97 (*Information Extraction: Multidisciplinary contributions to an emerging Information Technology*, Pazienza M.T.(Ed), Lecture Notes in Artificial Intelligence 1299, Springer-Verlag, Berlin Heidelberg New York, 1997) and what was debated at SCIE99 (and gathered in this book), as the current stage of the research and development in IE technology, the strong requirement for technology deployment emerges as a novelty, i.e. *the availability of robust adaptable systems to test either different methodologies or new application scenario without being forced to redefine knowledge resources and the kind of processing.* The first phase aimed at defining topics to be covered, at different extents of generality, in an IE system appears to be concluded; a new spirit calls for technological deployment for effective, adaptable IE systems!

I would like to thank individually all my colleagues from the Artificial Intelligence Research Group of the University of Roma Tor Vergata (and particularly Roberto Basili and Michele Vindigni) who supported my efforts at organizing SCIE99 and editing this book.

Roma, July 1999 Maria Teresa Pazienza

Organization

SCIE99 is organized by the University of Roma, Tor Vergata (Italy).

Program Committee

Luigia Carlucci Aiello (University of Roma "*La Sapienza*")
Elisa Bertino (University of Milano)
Domenico Sacca' (University of Calabria)
Lorenza Saitta (University of Torino)
Maria Teresa Pazienza (University of Roma, Tor Vergata)

Organizing Committee

Roberto Basili (University of Roma, Tor Vergata)
Cristina Cardani (University of Roma, Tor Vergata)
Maria Teresa Pazienza (University of Roma, Tor Vergata)
Michele Vindigni (University of Roma, Tor Vergata)
Fabio Massimo Zanzotto (University of Roma, Tor Vergata)

Supporting Institutions

The SCIE99 has been partially supported by

- AI*IA, Italian Association for Artificial Intelligence
- ENEA, National Institute for Alternative Forms of Energy
- ESA, European Space Agency
- University of Roma, Tor Vergata, Italy

Table of Contents