Lecture Notes in Artificial Intelligence 2700

Edited by J. G. Carbonell and J. Siekmann

Subseries of Lecture Notes in Computer Science

Springer
Berlin
Heidelberg
New York
Hong Kong
London
Milan
Paris
Tokyo

Maria Teresa Pazienza (Ed.)

Information Extraction in the Web Era

Natural Language Communication for Knowledge Acquisition and Intelligent Information Agents

Springer

Series Editors

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

Volume Editor

Maria Teresa Pazienza
AI Research Group
Department of Computer Science, Systems and Production
Via del Politecnico 1
00133 Roma, Italy
E-mail: pazienza@info.uniroma2.it

Cataloging-in-Publication Data applied for

A catalog record for this book is available from the Library of Congress.

Bibliographic information published by Die Deutsche Bibliothek
Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data is available in the Internet at <http://dnb.ddb.de>.

CR Subject Classification (1998): I.2, H.3, H.2.8, H.4

ISSN 0302-9743
ISBN 3-540-40579-8 Springer-Verlag Berlin Heidelberg New York

Springer-Verlag Berlin Heidelberg New York
a member of BertelsmannSpringer Science+Business Media GmbH

http://www.springer.de

Printed in Germany

Typesetting: Camera-ready by author, data conversion by PTP-Berlin GmbH
Printed on acid-free paper SPIN: 10928653 06/3142 5 4 3 2 1 0

Preface

The number of research topics covered in recent approaches to Information Extraction (IE) is continually growing as new facts are being considered. In fact, while the user's interest in extracting information from texts deals mainly with the success of the entire process of locating, in document collections, facts of interest, the process itself is dependent on several constraints (e.g. the domain, the collection dimension and location, and the document type) and currently it tackles composite scenarios, including free texts, semi- and structured texts such as Web pages, e-mails, etc.

The handling of all these factors is tightly related to the continued evolution of the underlying technologies.

In the last few years, in real-world applications we have seen the need for scalable, adaptable IE systems (see M.T. Pazienza, "Information Extraction: Towards Scalable Adaptable Systems", LNAI 1714) to limit the need for human intervention in the customization process and portability of the IE application to new domains. Scalability and adaptability requirements are still valid impacting features and get more relevance into a Web scenario, wherein intelligent information agents are expected to automatically gather information from heterogeneous sources.

In such an environment, the ability to manage different kinds of knowledge assumes an important role. As the problem of knowledge acquisition cannot be solved only by human intervention (with the increased dimension and distribution of collections it becomes too costly and time consuming), the process of automatic knowledge acquisition becomes crucial to scale up the systems. See the contribution by R. Yangarber, "Acquisition of Domain Knowledge":

"Linguistic knowledge in Natural Language understanding systems is commonly stratified across several levels. This is true of Information Extraction as well. Typical state-of-the-art Information Extraction systems require syntactic-semantic patterns for locating facts or events in text; domain-specific word or concept classes for semantic generalization; and a specialized lexicon of terms that may not be found in general-purpose dictionaries, among other kinds of knowledge.

The objective of IE, as considered here, is to analyze text written in plain natural language, and to find facts or events in the text. The facts and events are formally expressed as multi-argument (n-ary) relations, whose arguments are entities, corresponding to objects in the real world. Information Extraction systems typically operate within a specific domain, and need to be adapted for every new domain of interest. Adaptation for a particular domain entails the collection of knowledge that is needed to operate within that domain.

There has been an observable trend in these approaches, moving from the labor-intensive manual methods of customization, toward automatic methods of knowledge acquisition; further, among the automatic methods, moving from fully-

supervised methods, which require large amounts of annotation, toward unsupervised or minimally supervised methods."

A further support to speed up the localization of *facts-of-interest* in texts descends from *terminology awareness*. There is a demand for reliable methods both to identify key terms or phrases characterizing texts and to link them with other texts and knowledge sources. See the contribution by B. Daille, "Terminology Mining":

"Terminology mining is a major step forward in terminology extraction and covers acquisition and structuring of the candidate terms.

The computation methods used for terminology mining depend on the data to be processed: raw texts or texts enhanced with linguistic annotations, and the use or not of pre-existing knowledge sources.

In terminology mining, references are made to the acquisition of complex terms, the discovering of new terms, but also the structuring of the acquired candidate terms. Among others, it is possible to adopt lexicality, criteria of the linguistic well-formedness of complex terms and their variations expressed in terms of syntactic structures.

We need to underline, for terminology extraction purposes, the crucial part of the handling of term variations in building a linguistic structuring, detecting advanced lexicalisation and obtaining an optimised representatives of the candidate term occurrences. Then we must analyse the implemented computational methods: shallow parsing, morphological analysis, morphological rule learning and lexical statistics."

Due to the fact *terms* are tightly connected to *concepts*, the latter plays a relevant role in characterizing the content of texts and, consequently, in relating them with other information sources. We have to focus on the aspect of the representativeness of a term to provide a mathematically sound formulation of its relatedness to concepts. See the contribution by T. Hisamotsu and J. Tsujii, "Measuring Term Representativeness":

"Although conventional methods based on tf-idf and its variants used intensively in IR systems have also been used to identify terms in texts on the network, the empirical nature of such measures suggests that we should not use them in far more dynamic and heterogeneous situations such as those possible on the network.

Unlike conventional IR systems, we deal with many diverse text types with different lengths and subject fields, and we can not rely on the carefully calibrated parameters that the performances of these empirical measures are highly dependent on.

In order to understand why some empirical measures work well in certain applications but perform rather poorly in other application environments, we first have to disentangle the integrated nature of these measures and identify a set of new measures whose mathematical properties can be understood. Since the termhood comprises several different dimensions, each dimension should be understood in terms of its mathematical properties, and then different measures

that represent different aspects of termhood can be combined to accomplish goals specific to given application environments."

Information agents are emerging as a very important approach for building next generation value-added services. Among other uses, information agents could be profitable for automatically gathering information from heterogeneous sources. See the contribution by N. Kushmerick, "Finite-State Approaches to Web Information Extraction":

"It is possible to view information extraction as a core enabling technology for a variety of information agents. We therefore focus specifically on information extraction, rather than tangential (albeit important) issues, such as how agents can discover relevant sources or verify the authenticity of the retrieved content, or caching policies that minimize communication while ensuring freshness.

Scalability is the key challenge to automatic information extraction. There are two relevant dimensions. The first dimension is the ability to rapidly process large document collections. IE systems generally scale well in this regard because they rely on simple shallow extraction rules, rather than sophisticated (and therefore slow) natural language processing. The second and more problematic dimension is the number of distinct sources.

IE is challenging in this scenario because each source might format its content differently, and therefore each source could require a customized set of extraction rules. Machine learning is the only domain-independent approach to scaling along this second dimension. The use of machine learning could enable adaptive information extraction systems that automatically learn extraction rules from training data in order to scale with the number of sources."

To be successful, Information Agents should also be able to deal with linguistic problems. *Linguistic knowledge* is weighted to fruitfully support *intelligent agents* in the activities of filtering, selecting, and classifying a large amount of information daily available on the Web. Due to the unavailability of generic ontologies, intelligent agents behaviour is far from being semantically based. Meanwhile, it appears evident that it is important to take into account semantic aspects to obtain more precise results for IE systems.

Intelligent agents involved in the extraction process could be helpful in the process of mediation among several (possibly domain-specific) ontologies underlying different texts from which information could be extracted. See the contribution by M.T. Pazienza and M. Vindigni, "Agent Based Ontological Mediation in IE Systems":

"Different components should be considered when dealing with documents content. In fact different levels of problems arise related to the two existing communication layers: lexical and conceptual. A group of interacting agents could be seen as a small cooperative society requiring a shared language and communication channels in order to circulate ideas, knowledge and background assumptions. General abilities for language processing may be considered as part of the agent knowledge: differences in formal representation may be overcome by means of transposition – conversion mechanisms.

In the context of knowledge sharing we can refer to ontology as the means for specifying a conceptualisation. That is an ontology may be a description of concepts and relationships that can exist for an agent community for the purpose of enabling knowledge sharing and reuse, thus supporting ontological commitments (e.g. an agreement to use a vocabulary to put queries and make assertions in a way that is consistent – but not complete – with respect to the underlying theory).”

As a consequence of the easier access to textual information, there is a wider interest in not being limited to extract information in the context of a predefined application domain. *Open domain questioning* fascinates the new frontier of research in information extraction. See the contribution by D. Moldovan, ”On the Role of the Information Retrieval and Information Extraction in Question Answering Systems”:

“Question Answering, the process of extracting answers to natural language questions is profoundly different from Information Retrieval (IR) or Information Extraction (IE). IR systems allow us to locate relevant documents that relate to a query, but do not specify exactly where the answers are. In IR, the documents of interest are fetched by matching query keywords to the index of the document collection. By contrast, IE systems extract the information of interest provided the domain of extraction is well defined. In IE systems, the information of interest is in the form of slot fillers of some predefined templates. The QA technology takes both IR and IE a step further, and provides specific and brief answers to open domain questions formulated naturally.”

In a future agent-based adaptive Web information Extraction framework, possibly dialoguing with the user, we could think of *virtual agents* with linguistic abilities for interaction purposes. See the contribution by M. Cavazza, “Natural Language Communication with Virtual Actors”:

“The development of realistic virtual actors in many applications, from user interfaces to computer entertainment, creates expectations on the intelligence of these actors including their ability to understand natural language. Specific technical aspects in the development of language-enabled actors could be highlighted. The embodied nature of virtual agents leads to specific syntactic constructs that are not unlike sublanguages: these can be used to specify the parsing component of a natural language interface. However, the most specific aspects of interacting with virtual actors consist in mapping the semantic content of users’ input to the mechanisms that support agents’ behaviours. A generalisation of speech acts can provide principles for this integration.

Virtual agents are embodied in a physical (although virtual) environment: apart from the properties of any specific task they have to carry, this embodiment is at the heart of understanding the requirements for NLP. The embodiment of virtual agents requires that their understanding of language is entirely translated into actions in their environment.”

Hereafter, throughout the various sections, all the previously cited research topics will be dealt with and deeply analyzed. The papers represent the authors’ contribution to SCIE 2002, the “Summer Convention on Information Extrac-

tion," held in Frascati (Rome, Italy), in July 2002, attended by a very qualified international audience that participated actively to the technical discussions. Comments from the participants have been considered in some cases during updating and some further details were introduced.

It emerges that new technological scenarios are forcing IE research activities to move in new directions. Meanwhile question answering is being proposed as a new frontier. We shall see ...

June 2003

Maria Teresa Pazienza
Program Chair
SCIE 2002

Organization

SCIE 2002 – Summer Convention on Information Extraction – was organized by the AI research group (Maria Teresa Pazienza, Roberto Basili, Michele Vindigni and Fabio Zanzotto among others) of the University of Roma Tor Vergata (Italy), and was hosted by ESA – the European Space Agency – at the ESRIN establishment, its premises in Frascati, Italy. Special thanks to the staff of the European Space Agency (ESA) at ESRIN for valuable support in the organization of SCIE 2002.

Sponsoring Institutions

SCIE 2002 was partially supported by the following institutions:

AI*IA, Artificial Intelligence Italian Association, Italy
ESA, European Space Agency
DISP, Department of Computer Science, Systems and Production, University of Roma Tor Vergata, Italy
MIUR Ministero dell'Istruzione, dell'Universitá e della Ricerca, Italy
ENEA, Ente Nazionale Energie Alternative, Italy
NOUS Informatica srl., Italy

Table of Contents

Information Extraction in the Web Era