ArabOnto: experimenting a new distributional approach for building Arabic ontological resources

Ibrahim Bounhas*
Faculty of Sciences of Tunis,
Department of Computer Science,
University of Tunis,
1060 Tunis, Tunisia
E-mail: Bounhas.ibrahim@yahoo.fr
*Corresponding author

Bilel Elayeb
RIADI-GDL Research Laboratory,
The National School of Computer Sciences (ENSI),
2010 Manouba, Tunisia
E-mail: Bilel.Elayeb@riadi.rnu.tn

Fabrice Evrard
The Computer Science Research Institute of Toulouse (IRIT),
02 Rue Camichel, 31071, Toulouse, France
E-mail: Fabrice.Evrard@enseeiht.fr

Yahya Slimani
Faculty of Sciences of Tunis,
Department of Computer Science,
University of Tunis, 1060, Tunis, Tunisia
E-mail: yahya.slimani@fst.rnu.tn

Abstract: Ontologies are useful for modelling and retrieving knowledge in complex information systems. Ontology construction environments use statistical and linguistic information to extract knowledge from corpora. Within the great improvement in this field, there is a need to introduce the Arabic language in these environments. We present the ArabOnto architecture modelling the process of Arabic ontology extraction from corpora. ArabOnto focuses on linguistic issues related to Arabic term extraction and linking (i.e. from morphosyntactic parsing to clustering). We experiment our system by testing several alternatives on three domains. Besides, our ontologies are validated in the context of an information retrieval system.

Keywords: Arabic language; ontology development; distributional analysis; terminology organisation; semantic similarity; information retrieval; ontology evaluation.

Reference to this paper should be made as follows: Bounhas, I., Elayeb, B., Evrard, F. and Slimani, Y. (2011) ‘ArabOnto: experimenting a new distributional approach for building Arabic ontological resources’, Int. J. Metadata, Semantics and Ontologies, Vol. 6, No. 2, pp.81–95.

Biographical notes: Ibrahim Bounhas obtained a License Degree in Computer Science Applied to Management (2004) from the High School of Management of Tunis (ISG) and his Master’s Degree in Computer Science (2006) from the National School of Computer Sciences (ENSI). He is presently a PhD student at the Department of Computer Science of the Faculty of Sciences of Tunis (2007–2008). He presented a master thesis entitled “Un analyseur de contenu des documents scientifiques du web”. His current research interests are: ontology engineering, document analysis, Arabic text processing.

Bilel Elayeb is an Assistant Professor at the National School of Computer Science of La Manouba in Tunisia. He obtained his PhD in Computer Science from the National Polytechnic Institute of Toulouse and the National School of Computer Science, Tunisia in 2009. He obtained
1 Introduction

Within the increasing size of information systems and the generalisation of the web, there is a growing need for information technologies allowing capture and representation of knowledge, and help in accessing documents. In this context, terminological resources of several types are required (Bourigault and Lame, 2002). Any of these resources represents knowledge as a set of units (terms or concepts) structured through various types of relations. Different types of terminological resources share a common axis which organises units according to their similarities and differences (Charlet et al., 2006). In this area, graph-based algorithms are widely used (Mihalcea and Radev, 2006). Our work is based on distributional analysis (Cohen and Widdows, 2009; Pantel et al., 2009) used, for example, in Upery (Bourigault, 2002). Bourigault and Lame (2002) used a network of simple and compound Noun Phrases (NP) generated by a syntactic parser. They proposed to enrich it by distributional links. Then, it is exploited as a thematic index to browse a collection of documents and as a resource for a query expansion module. Bourigault and Lame (2002) call such a structure a ‘documentary ontology’. In specific domain corpora, this kind of resource constitutes a domain model which helps understand the domain structure and reduces the cognitive charge of the user while exploring and searching for documents. The context of our work in this paper consists in building Arabic ontologies structured with distributional links. We propose a new extension of the distributional method defined by Harris (1968) and extended by Bourigault (2002). In the remaining of this paper, we follow Bourigault and Lame (2002) by calling these resources ‘ontologies’ or ‘ontological resources’.

Building ontological resources is a process composed of many stages where several decisions should be taken. The main questions are: how to select relevant terms? And how to organise these terms in a coherent manner? Finally, the obtained knowledge should be visualised to provide more interactivity and understandability. Many researchers investigated and continue to work on acquiring domain models from textual and semi-structured corpora (e.g., books, scientific papers and encyclopaedias). Previous research has focused on co-occurrence and linguistic parsing. Building ontologies requires sophisticated corpus analysis tools. These tools are available for many languages (e.g., French and English). However, to the best of our knowledge, none of the existing approaches for Arabic document processing considered the ontology building process in all its steps in a domain-independent manner. Indeed, the main effort of researchers in Arabic NLP has been focused on the morphosyntactic level, thus developing linguistic resources of several types. We can refer to morphologic analysers (e.g., Buckwalter’s tool\textsuperscript{1}), knowledge bases (e.g., DIINAR\textsuperscript{2}) and syntactic parsers (e.g., Ditters and Koster, 2004). Besides, a great effort has been made to produce monolingual and multilingual corpora available, for example, through LDC\textsuperscript{3} and the NEMLAR/MEDAR projects.\textsuperscript{4} We can cite the Prague Arabic Dependency Treebank project, which contains texts annotated at the morphologic and syntactic levels (Smrz et al., 2007).

Despite this development, many challenges are still faced with knowledge extraction engines supposed to deal with Arabic texts. On the one hand, many of the existing tools for morphologic analysis and POS tagging are unable to provide a full parsing of these texts and in the same time