Master in Artificial Intelligence (UPC-URV-UB)

Master's Thesis Proposal

General Information

Master's Thesis Title:	Ontology-based information extraction
Publication Date:	March 2011
Expiry Date:	October 2011
Modality:	☐ technological project ☒ research work
M.Sc. Th. Advisor:	Dr Antonio Moreno, Dr David Sánchez
M.Sc. Th. Advisor's Dept. & University:	Computer Science and Maths. Dep., Univ. Rovira i Virgili
M.Sc. Th. Advisor e-mail:	{antonio.moreno, david.sanchez}@urv.cat
Observations:	Work to be developed at URV in Tarragona
Student's Name: (if already known)	Carlos Vicient Monllaó

M.Sc. Thesis Description

Main issues / Brief Description:

One of the basic pillars of the Semantic Web concept is the idea of having explicit semantic information on the Web pages, that can be used by intelligent agents in order to solve complex problems of Information Retrieval and Question Answering. It is important to develop mechanisms that allow the automatic annotation of Web pages, since it is not reasonable to expect users to perform this process manually. The aim of this work is to design, implement and evaluate new techniques that allow to annotate automatically the content of a textual/semi-structures/structured information source, taking into account the semantic information given by a domain ontology and other knowledge repositories such as WordNet.

Detailed Description:

Thanks to the increasing amount of freely available electronic textual data, there has been a growing interest in the research community in developing data mining techniques which are able to exploit this kind of information. However, textual documents describing a particular entity (e.g. questionnaires, Wikipedia entries, etc.) are difficult to process in order to extract relevant features which could be exploited in order to apply semantically focused data mining algorithms [Hotho,04].

Semantic-based information extraction relies on ontologies in order to interpret the textual content of a resource regardless of its format. Even though there have been many conceptual approximations in the field of Semantic Web in which it is assumed that resources have been semantically annotated, in the short-term future we cannot expect the availability of a massive amount of annotated Web resources. So, in order to take profit from the Web resources which are currently available, the extraction of features from plain text, as it is proposed in this work, goes through the syntactic analysis of its content and its association with the concepts modelled in one or more input ontologies.

With respect to the feature extraction problem from structured and semi-structured resources, they rely on methods of the *Information Extraction* field. They are typically based on the fact that the document presents a certain structure (e.g. attribute-value pairs) in order to correctly interpret its content. The information extraction process is implemented by means of a series of rules which depend on the document structure (*wrappers*). Even though this approximation may be useful in some cases (for example, web sites about products), it has the inconvenient of the lack of semantic analysis and its total dependency on the document format.

The main aim of this work is to design and implement a novel method that is able to extract relevant features from a range of textual documents going from complete plain textual data to semi-structured and structured resources. The designed methodology should be able to take profit from pre-processed input when it is available in order to complement its own learning algorithms. The key point of the work is to complement the syntactical parsing and several natural language processing techniques with the knowledge contained in one or several input ontologies in order to be able to 1) identify relevant features describing a particular entity from textual data, 2) to associate, if applicable, extracted features to concepts contained in the input ontologies. In this manner, the output of the system would consist on tagged features which can be directly exploited by semantically grounded data mining algorithms (e.g. clustering) in order to classify them.

The main tasks of the work are:

1) To compile a state of the art in ontology-based and ontology-driven information extraction systems, paying also especial attention to related fields such as automatic semantic annotation of documents [Cimiano05]. As a result, the main techniques and methodologies used by related works should be identified and analysed.
2) To design a methodology that, taking raw text describing a certain entity as input would be able to:
    a. Detect features describing or associated to the entity. This stage will focus on Named Entities.
    b. To assess which of the extracted features are more closely related to the entity (i.e. they better identify and describe it) in order to maximize the accuracy of the data mining process.
    c. To associate selected features to concepts modelled in one or several input ontologies, if they fit in the domain covered by the ontology.
3) To study how to modify/adapt the designed methodology to take profit from semi-structured and structured resources (e.g. pre-tagged Wikipedia articles) in order to assist the automatic learning.
4) To carefully evaluate the methodology as a whole and each learning stage individually, comparing the accuracy regarding the employed techniques, the involved parameters and the type of input.

The designed methodology should be as automatic, unsupervised and domain independent as possible in order to maximize its generality and applicability [Sánchez08]. Natural anguage related problems such as ambiguity should be considered in order to improve the quality of the results. The scalability of the approach should be also carefully considered, minimizing the dependency on external resources. Unsupervised learning techniques such as the use of statistical analyses evaluating information distribution [Turney01] and general linguistic patterns [Etzioni05] can aid on this purpose.

References:

[Cimiano05] Cimiano, P., Ladwig, G., Staab, S.: Gimme' the context: context-driven automatic semantic annotation with c-pankow. Actas de 14th World Wide Web Conference (2005).

[Etzioni05] Etzioni, O., Cafarella, M., Downey, D., Popescu, A.M., Shaked, T., Soderland, S., Weld, D.S. and Yates, A.: Unsupervised named-entity extraction from the Web: An experimental study. Artificial Intelligence 165. 2005. 91-134.

[Hotho,04] A. Hotho, A. Maedche, S. Staab: Ontology-based Text Document Clustering. Künstliche Intelligenz 16(4), 48-54 2001.

[Sánchez08] Sánchez, D.: Domain Ontology Learning from the Web. Phd Thesis. VDM Verlag. 2008

[Turney01] Turney, P.D.: Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. Actas de 12th European Conference on Machine Learning. Freiburg, Germany 2001. 491-499.

Minimal Requirements & Previous Knowledge:

Ontologies, natural language parsing, semantic similarity measures, annotation techniques, linguistic patterns

Other comments:

The work is included in the funded research project DAMASK: Data Mining Algorithms with Semantic Knowledge (TIN2009-11005).

Location and Date: Tarragona, March 1st, 2011

To the Academic Commission of the Master in Artificial Intelligence (CAIMIA)