DEVELOPMENT OF AN ONTOLOGY CONSTRUCTION COMPONENT FOR THE OBCIE (ONTOLOGY-BASED COMPONENTS FOR INFORMATION EXTRACTION) APPROACH

D. C. Wimalasuriya[1], M. S. Lewke Bandara[2]

[1] Department of Computer Science and Engineering, Faculty of Engineering, University of Moratuwa, Sri Lanka.
Email: chinthana,128218h@uom.lk
[2] Department of Computer Science and Engineering, Sri Lanka Institute of Information Technology (SLIIT),
Malabe, Sri Lanka.

ABSTRACT

Information extraction systems identify and retrieve certain types of information from natural language text. A recent development in the field of information extraction is the emergence of ontology-based information extraction as a sub-filed, where ontologies are used to guide the information extraction process and to present the extracted information.

One of the challenges faced by fields of ontology-based information extraction and information extraction is the difficulty of reuse of prior work in developing new systems. A component-based approach for information extraction named OBCIE (Ontology-Based Components for Information Extraction) has been previously developed to address this issue. This paper presents the progress in developing an ontology construction component for the OBCIE approach, which identifies classes and relationships for a given domain. It is centered on discovering the information contained within the loose structure of Wikipedia pages.

Key words: information extraction, ontologies, components

1. INTRODUCTION

Information extraction is a sub-field of natural language processing (NLP) which aims to identify and retrieve or extract some information from natural language text. Ontology-based information extraction has recently emerged as a sub-field of information extraction. Here, the objective is to use ontologies to guide the information extraction process and to present the results. The concept of an ontology comes from the field of knowledge representation, where it is defined as a formal and explicit specification of a shared conceptualization [1].

Although a fairly large number of information extraction and ontology-based information extraction systems have been developed by researchers, their usage is not widespread or commercial. Lack of effective mechanisms for reuse has been identified as one major reason behind this. The first author has previously developed a component-based approach for information extraction named OBCIE (ontology-based components for information extraction) [2] that attempts to address this issue. It aims to derive the advantages of the use of software components in developing information extraction and is thus related to reuse-oriented software engineering. The salient features of this approach are as follows.

- Information extractors: components which make extractions with respect to particular ontological concepts.
- Platforms for information extraction: domain, concept and corpus independent implementations of information extraction techniques.
- A series of operations that describe how the system functions.
- Ontology construction: identifying classes and properties of the ontology.

Figure 1 presents the main components of the OBCIE approach and their interaction.

The authors' previous works have developed components for all the different functional areas defined by the OBCIE approach, except ontology construction. The requirement for this component is to produce an ontology for a given domain as a specification of the web ontology language (OWL) [3], which has emerged as a de facto standard in defining ontologies. The other components of the OBCIE approach make use of this OWL ontology in guiding the ontology population task, which aims to identify *instances* and *property values* fitting into the templates provided by the classes and properties of the constructed ontology. While several ontology-based information extraction systems that perform the task of ontology construction have been developed, the methodologies used by them have not been converted into components compatible with the OBCIE approach prior to the work presented here.

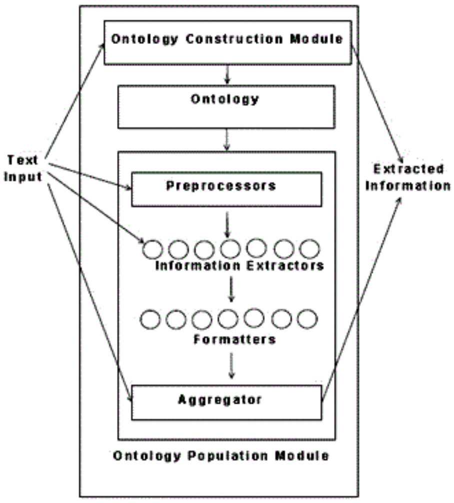


Figure 1: An ontology-based information extraction system under the OBCIE approach

The rest of the paper presents the progress in developing an ontology construction component compatible with the OBCIE approach.

2. METHODOLOGY

The ontology construction component for the OBCIE approach is being developed to leverage the information contained in Wikipedia. Wikipedia (**http://www.wikipedia.org**) is an online encyclopedia, which relies on the contributions made by collaborators on a volunteer basis. It is seen as one of the most successful crowd-sourcing applications. By January 2013, it has grown into over 24 million articles and contains pages on almost any topic.

In using Wikipedia for the purpose of ontology construction, we pay special attention to its following features.

1. Infoboxes: These are the boxes present in the top right-had corner of many Wikipedia pages. They present a summary of the information contained in the page. In the process, they identify *values* for a set of *properties* for the entities presented by the respective pages.

2. Categories: Wikipedia pages are categorized into a loose hierarchy of pages created by the contributors. It is not a strict hierarchy because new categories can be freely introduced by contributors and a page may belong to a number of categories. Still, it is quite a close to a hierarchy of classes that is included in an ontology.

Figure 2 shows an example for infoboxes and categories.

In performing ontology construction for the OBCIE approach, the focus is on developing *domain ontologies*, which provide a specification for some domain (e.g., higher educational institutes, terrorist attacks, consumer product reviews, etc.). Hence, the ontology construction component attempts to develop a domain ontology for a given domain using the information contained in Wikipedia, concentrating on infoboxes and categories.

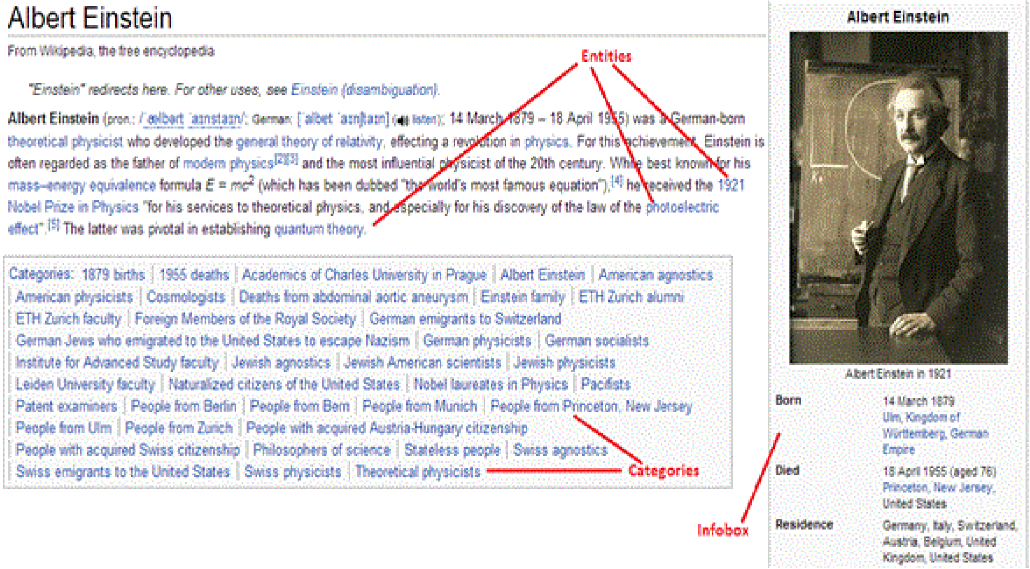
The following are the main steps followed by the component in developing an ontology for a given domain.

1. Obtain a set of *seed concepts* for the domain from a human. For example, for the domain of higher education institutes university, professor and degree can be used as seed concepts.
2. Find Wikipedia pages presenting information of instances of seed concepts.
3. Extract the properties and relationships of seed concepts using the structure of infoboxes and the category hierarchy of the Wikipedia pages discovered in Step 2.
4. Organize the results into an OWL ontology.
5. Attempt to improve the ontology through the following approaches.
    a. Discover related concepts from the WordNet lexical-semantic database for the English language [4].
    b. Discover new datatype and object properties from the Wikipedia pages themselves (as opposed to infoboxes and categories)
6. Present the revised ontology to humans, who may or may not be domain experts, through an ontology editor such as the Protégé tool (**http://protege.stanford.edu/**).

This approach is influenced by several works including the following.

- The Kylin Ontology Generator used by the Kylin system [5], which generates ontologies using a similar approach. However, the Kylin Ontology Generator does not use seed concepts or present the output ontology as an OWL ontology.

- Work by Hwang [6], who has used seed concepts in attempting to automatically develop ontologies back in 1999.



Albert Einstein

From Wikipedia, the free encyclopedia

"Einstein" redirects here. For other uses, see Einstein (disambiguation).

Albert Einstein (pron.: /ˈælbərt ˈaɪnstaɪn/; German: [ˈalbɐt ˈaɪnʃtaɪn] (listen); 14 March 1879 – 18 April 1955) was a German-born theoretical physicist who developed the general theory of relativity, effecting a revolution in physics. For this achievement, Einstein is often regarded as the father of modern physics[2][3] and the most influential physicist of the 20th century. While best known for his mass–energy equivalence formula $E = mc^2$ (which has been dubbed "the world's most famous equation"),[4] he received the 1921 Nobel Prize in Physics "for his services to theoretical physics, and especially for his discovery of the law of the photoelectric effect".[5] The latter was pivotal in establishing quantum theory.

Categories: 1879 births | 1955 deaths | Academics of Charles University in Prague | Albert Einstein | American agnostics | American physicists | Cosmologists | Deaths from abdominal aortic aneurysm | Einstein family | ETH Zurich alumni | ETH Zurich faculty | Foreign Members of the Royal Society | German emigrants to Switzerland | German Jews who emigrated to the United States to escape Nazism | German physicists | German socialists | Institute for Advanced Study faculty | Jewish agnostics | Jewish American scientists | Jewish physicists | Leiden University faculty | Naturalized citizens of the United States | Nobel laureates in Physics | Pacifists | Patent examiners | People from Berlin | People from Bern | People from Munich | People from Princeton, New Jersey | People from Ulm | People from Zurich | People with acquired Austria-Hungary citizenship | People with acquired Swiss citizenship | Philosophers of science | Stateless people | Swiss agnostics | Swiss emigrants to the United States | Swiss physicists | Theoretical physicists

Albert Einstein

Albert Einstein in 1921

Born	14 March 1879 Ulm, Kingdom of Württemberg, German Empire
Died	18 April 1955 (aged 76) Princeton, New Jersey, United States
Residence	Germany, Italy, Switzerland, Austria, Belgium, United Kingdom, United States

Figure 2: The infobox and categories for a Wikipedia page (http://en.wikipedia.org/wiki/Albert_einstein accessed on February 17, 2013)

3. RESULTS

This is an ongoing project and we are not in a position to present a final set of results. Still the preliminary results obtained so far provide enough evidence to justify the viability of the approach. The important results obtained so far are as follows.

- A set of seed concepts have been identified for the domains of *higher education institutes*, *commercial agriculture* and *elections*. These domains are expected to be used in testing the system being developed.
- The task of searching Wikipedia and finding the pages relevant to seed concepts is close to completion.
- Work is currently being carried out on developing ontologies using infobox structures and category hierarchies. Obstacles faced in this task include duplicate elimination and removing unrelated properties and relationships.
- Some progress has been made in converting the ontology being developed into OWL format. The use of Jena Ontology API (**http://jena.apache.org/documentation/ontology/**) is being investigated here.
- Supplementing the ontology with information from Wikipedia pages and the WordNet has not been started yet.

4. CONCLUSION

As of now, the development of the ontology construction component compatible with the OBCIE approach is still work-in-progress. Once fully developed, it can be integrated with the other components of the approach giving rise to a system that can perform ontology-based information extraction for a given domain with minimal input – in the form of seed concepts and manual review of the constructed ontology. This presents a significant improvement over relying on a human for the development of an ontology.

5. REFERENCES

[1] T. Gruber, "*Ontolingua : a translation approach to providing portable ontology specifications*", Knowledge Acquisition, 5(2), pp. 199-220, 1993

[2] D. Wimalasuriya, D. Dou, "*Components for Information Extraction: Ontology-Based Information Extractors and Generic Platforms*", in: Proceedings of the 19th ACM Conference on Information and Knowledge Management, pp. 9-18, 2010

[3] G. Antoniou, F. van Harmelen, "*Web ontology language: OWL", Handbook on Ontologies, pp.*

67-92, 2004

[4] C. Fellbaum (Eds.), "*WordNet: An Electronic Lexical Database",* MIT Press, 1998.

[5] F. Wu, D.S. Weld, "*Autonomously semantifying Wikipedia*", in: Proceeding of the 16th ACM Conference on Information and Knowledge Management, pp. 41-50, 2007

[6] C.H. Hwang, "*Incompletely and imprecisely speaking: using dynamic ontologies for representing and retrieving information*", in: Proceedings of the 6th International Workshop on Knowledge Representation meets Databases, pp. 14-20, 1999.