A refined model of ontology-driven information extraction

Chunyu Cong[1, a], Shan Huo[1] , Xiao Meng[2, b*], Rui Gao[3] and Zhongying Wang[4]

[1] Changchun University of Chinese Medicine, Changchun 130117, China;

[2]Chinatelcom Jilin Corporation, Changchun 130000, China.

[3]Jilin Provincial Institute of Education, Changchun 130000, China

[4]Continental Automotive Changchun Co.,Ltd , Changchun 130033, China

[a]644497215@qq.com [b]18904300295@189.com

Keywords: Ontology, information extraction

Abstract. An ontology is a formal and normalized explanation of a shared conceptualization while information extraction (IE) is a form of natural language processing in which certain types of information must be recognized and extracted from text. The methods of ontology-based IE fall in two broad categories: document-driven IE and ontology-driven IE. Document-driven IE is known as semantic annotation which annotates and manages the knowledge in semantic web with the semantic information in domain ontologies. Ontology-driven IE can extract information from unstructured documents based on a domain ontology. In this paper, we use ontology-driven IE to extract hazard information from Chinese food complaint documents and the results are delightful.

Introduction

An ontology is an explicit specification of a conceptualization. The term is borrowed from philosophy, where an Ontology is a systematic account of Existence. For AI systems, what "exists" is that which can be represented. When the knowledge of a domain is represented in a declarative formalism, the set of objects that can be represented is called the universe of discourse. This set of objects, and the describable relationships among them, are reflected in the representational vocabulary with which a knowledge-based program represents knowledge. Thus, in the context of AI, we can describe the ontology of a program by defining a set of representational terms. In such an ontology, definitions associate the names of entities in the universe of discourse (e.g., classes, relations, functions, or other objects) with human-readable text describing what the names mean, and formal axioms that constrain the interpretation and well-formed use of these terms. Formally, an ontology is the statement of a logical theory.

Information extraction (IE) is the task of automatically extracting structured information from unstructured and/or semi-structured machine-readable documents. In most of the cases this activity concerns processing human language texts by means of natural language processing (NLP). Recent activities in multimedia document processing like automatic annotation and content extraction out of images/audio/video could be seen as information extraction.

Ontology-based information extraction has recently emerged as a subfield of information extraction. Here, ontologies - which provide formal and explicit specifications of conceptualizations - play a crucial role in the information extraction process. An ontology is a formal and normalized explanation of a shared conceptualization while information extraction is a form of natural language processing in which certain types of information must be recognized and extracted from text[1-2]. Information extraction (IE) is the task of identifying, collecting and normalizing relevant information from natural language text and skipping irrelevant text passages. IE systems do not attempt an exhaustive deep natural language analysis of all aspects of a text. Rather, they are built in order to analyze or “understand” only those text passages that contain information relevant for the task at hand[3].

Ontology-driven IE

The model of ontology-driven IE is proposed by D.W.Embley for the first time[4-7]. It can extract information from unstructured documents based on a domain ontology. The input of the system is domain ontology and unstructured documents. The information extraction system (IES) proposed by D.W.Embley is automatic, portable and resilient to changes in source-document formats. The only step that requires significant human intervention is the initial creation of an application ontology. However, the system is high-efficiency only when the documents are data rich, which means the documents have a number of identifiable constants such as dates, and narrow in ontological breadth, which means the domain of interest can be described by a lightweight ontology. While, the domain which an ontology represents will change inevitably over time so it is important to keep the ontology up-to-date. The domain ontology will no longer change once it is taken as the input of the D.W.Embley's model. We need the IES provide services that are able to adapt the underlying ontology to the current reality. Burcu Yildiz and Silvia Miksch propose another model of IES[80]. In this module, an ontology management module is integrated into the IES that provides the functionalities of ontology learning and population, and change management of ontology. The model is shown in Fig. 1.

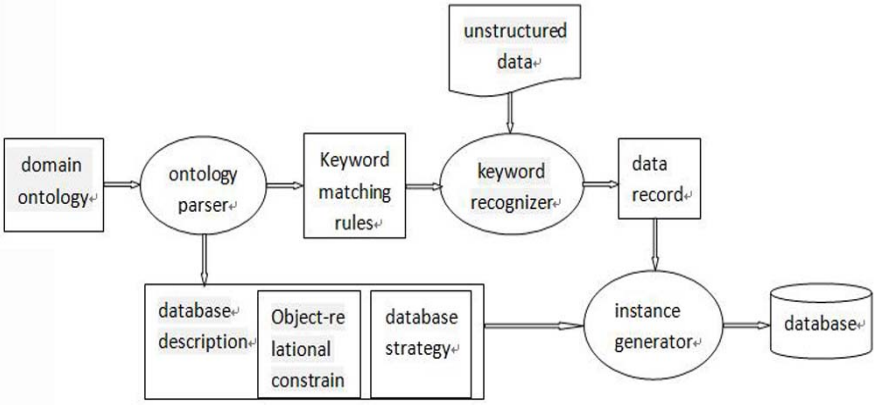

Fig. 1 The model of Ontology-driven IE

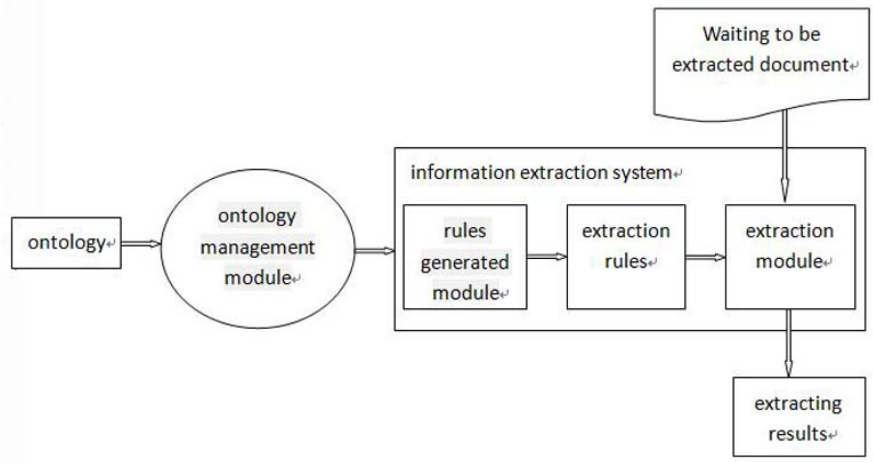

Fig. 2 Refined model of Ontology-driven IE

Refined Ontology-driven IE.

Although the IES proposed by Burcu Yildiz and Silvia Miksch can make up some shortcomings of D.W.Embley's IES, there are also some drawbacks. It is lack of semantic reasoning ability, and the extracted information can not make sense sometimes. That causes the problem of information fragment. Although the IES provides the functionalities of ontology learning and change management, these all require human intervention. There is not a method of ontology self-learning to

achieve the goal of ontology expansion in order to complete the task of information extraction more comprehensively and accurately. Moreover, it will cause the problem of semantic deviation when we use the imperfect ontology to extract the sparse information in short documents. The perfection of the ontology knowledge base is essential to the domain of ontology-based information extraction. We can extract information more accurately and that will reflect more semantic information in the documents. The refined model of ontology-driven IE is shown in Fig. 2.

Experiment Results

We construct an ontology as is shown in Fig. 3.

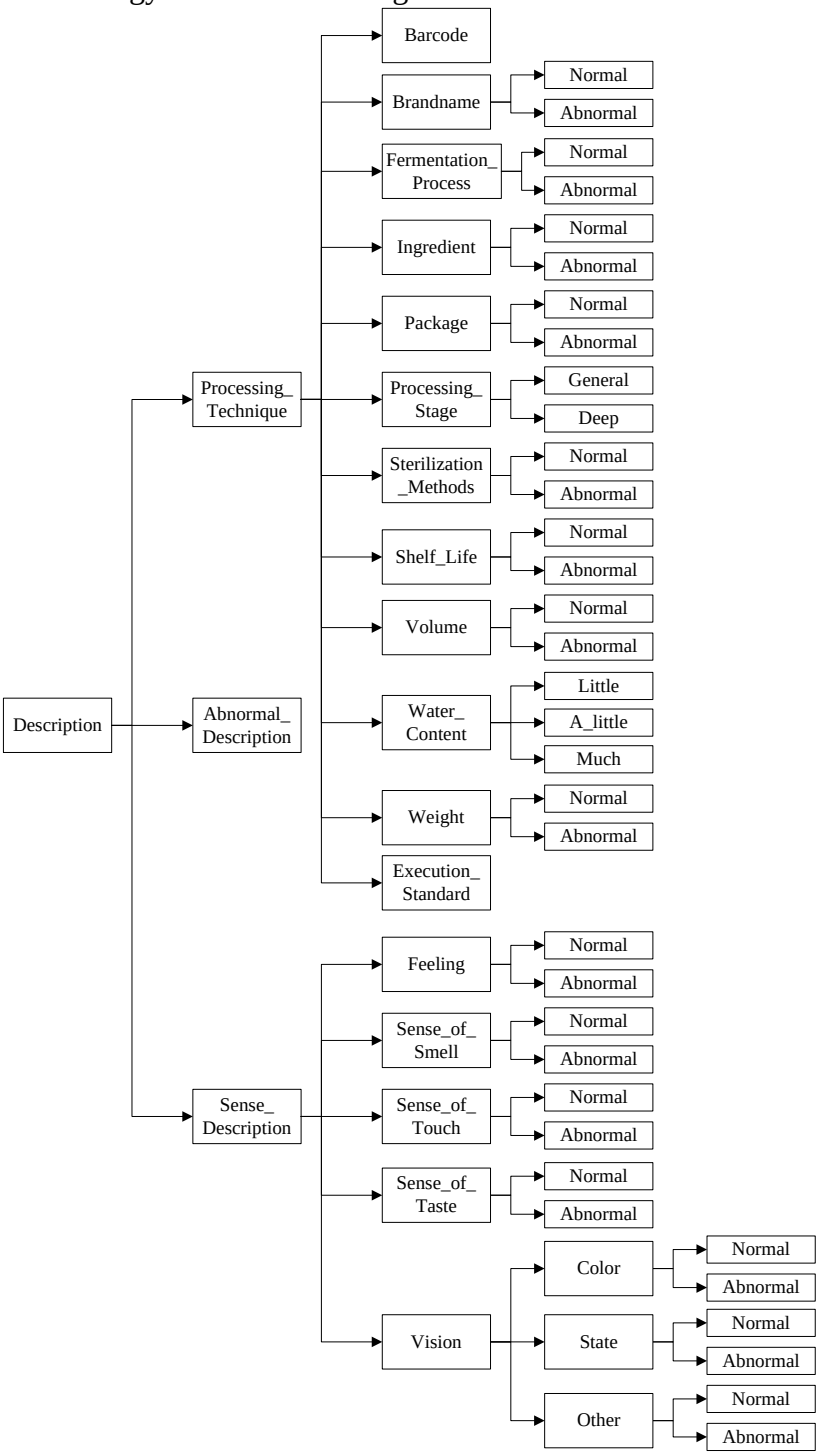

Fig. 3 The ontology model

This paper uses information evaluation index including the traditional measures of precision (P), recall (R) and F_measure to evaluate the results. The formula is shown below:

$$P = \frac{manually\_selected \cap machine\_selected}{machine\_selected}$$

$$R = \frac{manually\_selected \cap machine\_selected}{mannually\_selected}$$

With “manually_selected” denoting the set of food complaint documents whose hazard information was extracted by humans, “machine_selected” denoting the set of food complaint documents whose hazard information was extracted by our mehtod.

The weighted harmonic mean of precision and recall, F_measure is calculated as:

$$F\_measure = \frac{2 \times P \times R}{P + R}$$

We first preprocessed 1000 dairy complaint documents downloaded from the internet and the results are shown in Table 1.

Table 1 The results of refined ontology-driven IE

P	R	F_measure
95.47%	93.2%	94.32%

Summary

In this paper, we proposed a model of refined ontology-driven IE and the extraction results are delightful. However, we still have a lot of work to do. In the future we will expand our experiment to other domains and try to generate a new domain ontology based on our food ontology automatic or semi-automatic. The semantic reasoning ability of our method is limited. We will try to rectify our algorithms to improve the ability of semantic reasoning in the future. Although the granularity of our method is sentenced, we still extract words instead of sentence in the first place. So we still have to solve the problem of information fragment in the future.

References

[1]Studer R, Benjamins V R, Fensel D. Knowledge Engineering, Principles and Methods. In journal of Data and Knowledge Engineering 25 (122) (1998) 161-197.

[2]Riloff, E. Information Extraction as a Stepping Stone Toward Story Understanding. In Understanding Language Understanding: Computational models of Reading, edited by Ashwin Ram and Kenneth Moorman, The MIT Press. 2002 435-460.

[3] Bootstrapping an Ontology-based Information Extraction System

[4] David W.Embley, Douglas M.Campbell, Randy D.Smith. Ontology-Based Extraction and Structuring of Information from Data-Rich Unstructured Documents. In proceedings of the1998 ACM 7th International Conference on Information and Knowledge Management 1998 52-59.

[5] David W.Embley, Douglas M.Campbell. A Conceptual-modeling approach to extracting data from the Web. In Proceedings of the 17th International Conference on Conceptual Modeling1998 78-91.

[6] David W. Embley , Douglas M. Campbell. Conceptual-model-based data extraction from multiple-record Web pages. In journal of Data & Knowledge Engineering 31(1999) 227-251.

[7] David W. Embley. Toward semantic understanding-An approach based on information extraction ontologies . In Proceedings of the Fifteenth Conference on Australasian Database Conference 2004 3-12.

[8] Burcu Yildiz, Silvia Miksch. Motivating ontology-driven information extraction. In proceedings of International Conference on Semantic Web and Digital Libraries 2007 45-53.