JOURNAL OF ENGINEERING annals OF FACULTY OF ENGINEERING HUNEDOARA
TOME VI (year 2008), FASCICULE 1, (ISSN 1584 – 2665)

AUTOMATIC TAXONOMY EXTRACTION USING DICTIONARY DATABASE

Boriana DELIISKA

DEPARTMENT OF COMPUTER SYSTEMS AND INFORMATICS, FACULTY OF MANAGEMENT, UNIVERSITY OF FORESTRY, SOFIA, BULGARIA

Abstract:
Ontologies are a central component of the Semantic Web infrastructure. However, the design and construction of ontologies is a labor-consuming and expensive process and requires allocation of huge resources in terms of cost and time. The main stage of domain ontology building is conceptualization or creating of taxonomy about the domain. In the article method and algorithm of automatic taxonomy extraction using dictionary database are investigated. In virtue of syntactic analysis of dictionary database taxonomy database is created including triples of subclass/superclass relationships between terms. On the base of the algorithm a software product TET is created. TET is applied for building taxonomy from existing dictionary database of computer science. The results and future research are discussed.
Keywords:
taxonomy, dictionary database, hierarchical relation, ontology

1. INTRODUCTION

The Semantic Web (SW) [12] has been proposed as an extension to the current Web where the content will be machineunderstandable. However, computers (or machines) today understand very little of available web content. Two resources necessary for realizing the semantic web are: (a) large scale availability of domain specific ontologies; and (b) large scale availability of annotations or metadata descriptions created by using terms, concepts or relationships provided by these ontologies [8].

Ontologies are a central component of the SW infrastructure. However, it is well acknowledged that design and construction of ontologies is a labor-consuming and expensive process and requires allocation of huge resources in terms of cost and time. The main stage of domain ontology building is conceptualization or creating of taxonomy about the domain. In the context of knowledge management, a taxonomy is a structured collection of terms, generally hierarchical, that is used for both classification and navigation [1]. Each term is linked by one or more hierarchical relations of type "parent/child" (*is-a/has-a*, *is-part-of/has-part* etc.) with the other terms. Manual taxonomy creating is often practice but very slow and subjective process.

Some researchers [11] have published relevant works, focusing on taxonomy extraction from text corpora.

On the other hand, an automatic taxonomy extraction from structured digital knowledge resources is convenient. Most digital dictionaries are fairly modest in

scope and are published as straight text, just like a print dictionary but more and more of them can be queried interactively as relational databases, and may offer other features.

At the present time there are a lot of dictionary databases (local or online) which content can be easy analyzed and reorganized for this purpose. The main goal of this research is mining of concept taxonomy from dictionary (lexical) database. It will be shown how taxonomy can be generated systematically from digital dictionary databases and how lexical resources are structured and administered and which declarative formalisms are usually used for this.

This paper is organized as follows. The method and algorithm of taxonomy extraction from dictionary database are described in Section 2. The application of the method and algorithm on computer science dictionary database is examined in Section 3. Section 4 discusses the conclusions and future work.

2. METHOD AND ALGORITHM OF TAXONOMY EXTRACTION FROM DICTIONARY DATABASE

The dictionaries contain structured lexical and other linguistic data, but it is not obvious how to extract these for use in artificial language processing systems. The classification of dictionaries is by the content organization and volume (glossary, lexicon, thesaurus, encyclopedia), by number of languages (monolingual, bilingual or multilingual dictionaries), by the domain (historical, technical, etc.), specialization (common and specialized), carrier (paper or electronic) target user (for human use and machine-readable), etc. [10].

For the purpose of taxonomy extraction the dictionaries of type lexicon are more convenient but any kind of dictionary can be used by underwritten method and algorithm. Lexicon (or vocabulary) is a book containing an alphabetical arrangement of the words in a language with their definitions and additional word-specific information.

Each dictionary includes syntactic, semantic and morphological information in different degree, which defines its structure. The dictionary database [7] has:

- megastructure, including an introductory part (a list of abbreviations and explanations), and final part (some appendixes);
- macrostructure containing the word list in alphabetical or systematical order consistent with the spelling of the entry words. This structure can be somewhat cyclic in nature with certain entries defined in terms of others;
- microstructure. The material presented in the dictionary is not connected in a consistent manner, but it is segregated into thousands of small chapters called dictionary entries. The structure of dictionary entry includes term (headword), definition, grammar and phonetic labels, context, links to related terms etc.;
- mesostructure which is not obligatory. It includes: a) Interrelations of lexicon entries (hyponyms and hypernyms, synonyms, antonyms etc.) b) relations to external information.

The relationship of data structures to lexical rules will especially have to be specified. The method of taxonomy extraction from dictionary database includes the following preliminary conditions and rules:

- Only terms and its grammatical categories (without definitions, context etc.) are investigated.
- The terms of grammatical categories *noun phrases* and *nouns* are designated as classes (concepts) of the derivative taxonomy. The other terms (verbs, adjectives, adverbs etc.) are ignored.

- The hierarchical relations linking taxonomy classes (concepts) are of common type *is-subclass-of/has-subclass* or *is-superclass-of/has-superclass*.

Let suppose that the extracted taxonomy consists of

$T = \{t_1, t_2, ..., t_n\}$ -

$W = \{w_1, w_2, ..., w_m\}$,

where T is set of terms found in the taxonomy database and W is set of component words composing the terms in T; *n* and *m* are the total numbers of terms and words.

The terms are one- two- or multiword. Generally, could be discussed two types of subclass/superclass relationships between them [8]. In the first case, new terms are created by adding modifiers to existing terms. If t1 and t2 are terms in the same semantic category and t1 is nested in t2, then t1 is the head of t2, and t2 is the modifier of t1 [11]. In other words, t2 is hyponym of t1 and t1 is hypernym of t2.

For example, the term *address access time* was created by adding the modifier *address* to its hypernym *access time*, i.e. a triple of type:

$$access\ time \xrightarrow{is\text{-}superclass\text{-}of} address\ access\ time \text{ or}$$

$$address\ access\ time \xrightarrow{is\text{-}subclass\text{-}of} access\ time$$

Note: In English, specific terms are commonly compounds of a generic level term and some modifiers [3].

In the second case, subclass/superclass relationships are created between existing terms without common words. For example, although *terminator* (synonym of *completion flag* and *terminator symbol*) is a hyponym of *flag* and simultaneously of *sign*, it shares no common words with its hypernyms. In this case additional external information is necessary.

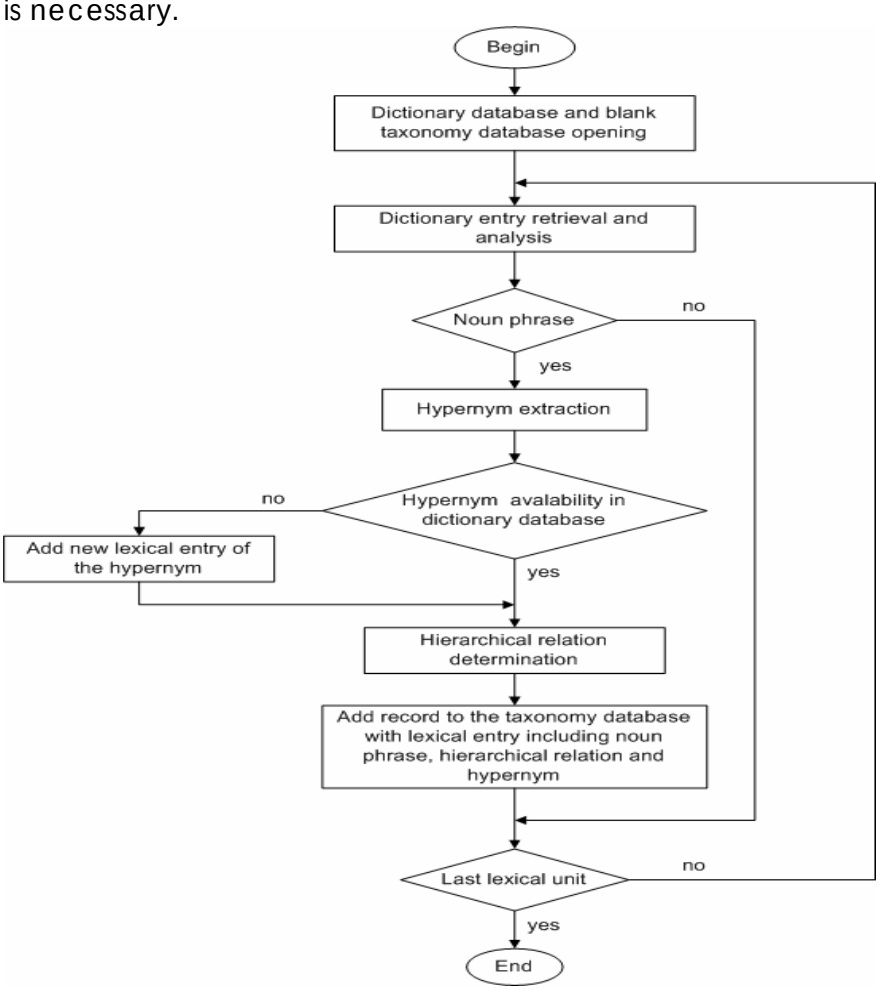

FIGURE 1. Algorithm for taxonomy extraction is created

In view of the above considerations an algorithm for taxonomy extraction is created (see fig.1). Accordingly this algorithm taxonomy database is generated from the macrostructure of dictionary database. Only dictionary terms of the first type of grammatical categories *noun phrase* and *noun* are analyzed. Noun phrase is decomposed and hypernym is extracted. In the first stage of this algorithm taxonomy entries representing triples of the above type are added to the taxonomy database.

In the second stage the taxonomy chains are deduced and generated in real time mode by multiple consecutive comparison of hypernyms and modifiers. These chains could be represented in graphical or text description. In the next stage the chain nodes with equal hypernyms are joined building the final taxonomy tree.

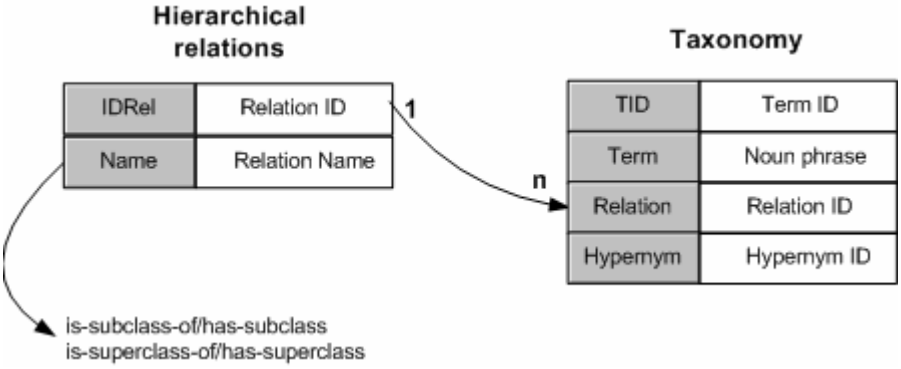

FIGURE 2. Schema of simple taxonomy database

The schema of simple taxonomy database on fig. 2 is shown. Each entry of the main table *Taxonomy* includes identifier (ID), noun phrase, relation ID and hypernym ID. The table *Relations* contains hierarchical relation types.

3. IMPLEMENTATION

On the base of proposed method and algorithm software is developed named TET (Taxonomy Extracting Tool). TET is applied in order to extract taxonomy from certain computer science dictionary databases [4, 6]. The databases are in MS Access format and are available in local and online variant.

For the present experiment the local dictionary database of computer science containing more of 10000 terms is used. Because of that TET was realized as Visual Basic for Application (VBA) module.

The syntactic parser of TET found that noun phrases and nouns in the dictionary database are about 74%. The rest are verbs, verb phrases, adjectives, adverbs and other grammatical categories. The generated taxonomy database contains about 7500 triples including noun phrases (classes) with its direct hypernyms. For example:

$$address\ latch \xrightarrow{is-subclass-of} latch$$
$$latch\ register \xrightarrow{is-subclass-of} register$$
$$register\ memory \xrightarrow{is-subclass-of} memory$$
$$memory\ block \xrightarrow{is-subclass-of} block, \text{ etc.}$$

The next step of TET is hypernym chain graph generation. For the above triples the following hypernym chain graph is deduced:

$$address\ latch \xrightarrow{is-subclass-of} latch \xrightarrow{is-subclass-of} register \xrightarrow{is-subclass-of} memory \xrightarrow{is-subclass-of} block \xrightarrow{is-subclass-of} \ldots$$

After that tree graph of the taxonomy is created joining equal nodes of all generated chains. Each hypernym is a node of the chain and on other hand can be regarded as a node of the tree graph (fig.3).

Obviously, there are irregular triples, for instance as the last triple of the above example. Really, the class *memory block* is superclass rather subclass of *block*. Moreover class *block* is not precised and because of that direct hyponyms illustrated on fig. 3 in fact refer to different kind of blocks (*hardware block*, *software block*, *data block* etc.).

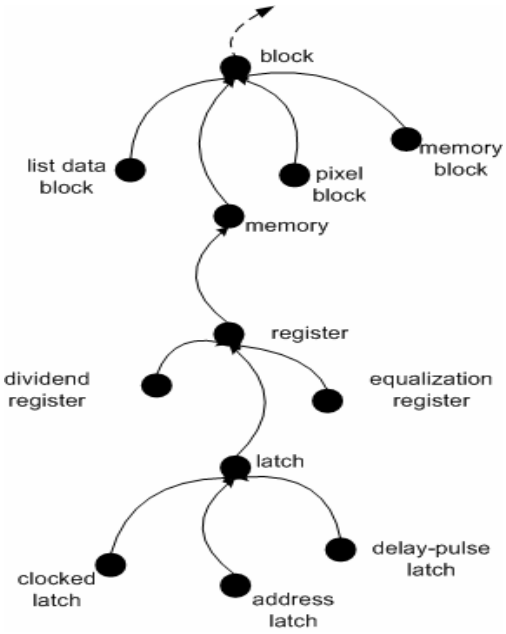

FIGURE 3. Direct hyponyms

The quantity of error triples depends on the precision of the syntactic parser in TET used for hypernym extraction. But such triples are possible because of incomplete terms in dictionary database.

Other kind of irregularities in the taxonomy tree (resp. in taxonomy database) are cycles, i.e. closed loops in chains. All irregularities have to be eliminated manually.

The precision of the automatically created taxonomy of computer science is about 65%. Some errors (about 14%) are due to the dictionary incompleteness. Therefore, the correctness of TET is about 50% in this concrete case.

As result of this experiment a taxonomy database is created as source for building simple ontology. On the other hand the taxonomy can be extended to thesaurus. According to [8], thesaurus is a taxonomy that also includes associated and related terms. Automatic extraction of associative and equivalence relations between classes in taxonomy database and other terms in dictionary database is very complex problem.

An algorithm of conversion thesaurus into ontology code is developed in [5] which could be applied for simple ontology code generation from taxonomy database.

4. CONCLUSIONS AND FUTURE WORK

The proposed method and algorithm are suitable for taxonomy extraction because save about the half of expert time necessary for simple domain ontology

building. Moreover the resultant taxonomy database in fact is a simple thesaurus which could be converted easily in ontology code.

The future efforts will be in direction to improvement of the syntactic parser with a view to increase correctness of generated triples. The choice of different dictionary databases in given domain and comparison of extracted taxonomies is another direction of future research in this area.

REFERENCES

[1.] Bailey S. Do you need a taxonomy strategy?, 2002 http://www.ikmagazine.com
[2.] Berners Lee T., Hendler J., Lassila O. The Semantic Web. Scientific American, May 2001
[3.] Croft W. Typology and Universals 2nd edition, Cambridge Textbooks in Linguistics, Cambridge Univ. Press, 2004
[4.] Deliiska B., Web-based system of dictionaries in area of computer science, Proceedings of Computer Science'2004, Sofia, 06-08.12.2004
[5.] Deliiska B., Thesaurus and domain ontology of geoinformatics, J. Transaction in GIS, ISSN: 1361-1682, vol. 12 issue 4 (11), 2007
[6.] Delijska B., P.Manoilov, Elsevier Dictionary of Computer Science, English-German-French-Russian, Elsevier, Amsterdam, ISBN 0-444-50339-0, 2001
[7.] Gibbon D., Modern Dictionary Making, http://www.spectrum.uni-bielefeld.de/Classes/Summer2003/Lexicography/lexicography/ lexicography.html, Version: July 24, 2003
[8.] Kashyap V., Ramakrishnan C., Thomas C., Sheth A., Taxa Miner: An Experimentation Framework for Automated Taxonomy Bootstrapping, International Journal of Web and Grid Services, Volume 1, Number 2 / 2005 p. 240-266
[9.] Lombardi V. 2003 Metadata Glossary, http://www.noisebetweenstations.com/personal/essays/metadata_glossary/metadata_glossary.html
[10.] Magee C.B.A., Computational Lexicography, http://www.cs.tcd.ie/courses/csll/magee00.ps
[11.] Pum-Mo Ryu, Key-Sun Choi, An Information-Theoretic Approach to Taxonomy Extraction for Ontology Learning, IOS Press, 2003
[12.] Semantic Web. W3C Semantic Web Activity. http://www.w3.org/2001/sw/