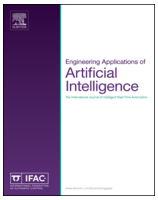
Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



Derivation of "*is a*" taxonomy from Wikipedia Category Graph

Mohamed Ben Aouicha [a,c,*], Mohamed Ali Hadj Taieb [b,c], Malek Ezzeddine [a,c]

[a] *Faculty of Sciences, Sfax University, Tunisia*
[b] *Higher Institute of Applied Sciences and Technology, Sousse University, Tunisia*
[c] *Multimedia InfoRmation system and Advanced Computing Laboratory, Sfax University, Sfax 3021, Tunisia*

ARTICLE INFO

Article history:
Received 18 July 2015
Received in revised form
18 December 2015
Accepted 18 January 2016

Keywords:
Taxonomy "is a"
Wikipedia Category Graph
DAG algorithms
Semantic similarity

ABSTRACT

Knowledge acquisition still represents one of the main challenging obstacles to designing intelligent systems exhibiting human-level performance in complex intelligent tasks. The recent developments in crowdsourcing technologies have opened new promising opportunities to overcome this problem by exploiting large amounts of machine readable knowledge to perform tasks requiring human intelligence. Wikipedia is a case of this research trend, being the largest collaborative and multilingual resource and linguistic knowledge that contains unstructured and semi-structured information. In this paper, we propose an approach for deriving "*is a*" taxonomy from the Wikipedia Categories Graph (WCG), which is an open collaborative resource. After building and filtering the WCG from a Wikipedia dump, the process would mainly consist in the exploitation of the "BY" tag and the sharing of plural headers. These methods provide a graph formed by a set of non-connected subgraphs. Therefore, we propose a process for linking them to finally obtain an "*is a*" taxonomy with only one root and modeled as a direct acyclic graph (DAG). In this work, specific DAG handling algorithms are used, including an algorithm for a DAG into sub-DAGs and another for merging two DAGs. The obtained taxonomy is assessed using semantic similarity measures, which consist in quantifying the likeness between two concepts or words. Therefore, we exploit a set of well-known benchmarks to compare the results obtained via the generated taxonomy to those achieved with WordNet, a resource created and maintained by domain experts. The experimental results revealed good correlations between computed values and human judgments. Compared to WordNet, the derived taxonomy was also noted to lead to an enhanced coverage capacity.

1. Introduction

Knowledge acquisition and modeling is a challenging long-standing problem that has been addressed in several ways. van Harmelen et al. (2007) provide a thorough survey on the various approaches so far proposed in the literature to deal with this problem. In fact, computer science researchers and specialists have been interested in structured knowledge for a long time. This knowledge has a significant role in various applications, such as Natural Language Processing, Artificial Intelligence, Information Retrieval and Knowledge Management. Gruber (1993) defines ontology as an example of knowledge structuring and as a mass of knowledge formally represented based on a conceptualization. Experts in ontology claim that the basic cognitive primitives represent concepts and the relations that link them. The need for a structured knowledge has led researchers to work on ontological knowledge resources and taxonomies. Taxonomy is a knowledge base organized in hierarchical way following only one kind of relations, such as "*is a*" or "*part of*". The "*is a*" taxonomy seeks to represent the generalization/specialization where ancestors are called "hypernyms" and descendants are called "hyponyms".

Manually built knowledge bases have several limitations. For example, WordNet[1] (Miller, 1995; Fellbaum, 1998) contains a few thousands of named entities and misses vocabulary of the specific fields, which negatively affects its coverage capacity. These limitations have recently prompted researchers to search for more efficient strategies, including the automatic extraction of various taxonomies (primarily "*is a*") starting from the collaborative encyclopedia Wikipedia.[2] The most notable advantage of Wikipedia is the broad range of topics it covers. In fact, several researchers have used it as the basis of their projects, including Cyc[3] (Lenat and Guha, 1989), DBpedia[4] (Auer et al., 2007), YAGO[5] (Hoffart et al., 2013), and BabelNet[6] (Navigli and Ponzetto, 2012).

The rest of the paper is organized as follows: Section 2 gives a brief overview of previous related works on the construction or derivation of

[*] Corresponding author.
E-mail address: mohamedali.hadjtaieb@gmail.com (M.A. Hadj Taieb).

[1] http://wordnetweb.princeton.edu/perl/webwn.
[2] http://en.wikipedia.org/wiki/Wikipedia.
[3] http://www.princeton.edu/~achaney/tmve/wiki100k/docs/Cyc.html.
[4] http://dbpedia.org/About.
[5] http://www.mpi-inf.mpg.de/~suchanek/downloads/yago/.
[6] http://babelnet.org/.

http://dx.doi.org/10.1016/j.engappai.2016.01.033
0952-1976/© 2016 Elsevier Ltd. All rights reserved.

ontologies and taxonomies based on Wikipedia. Section 3 presents the rich factors in the Wikipedia Categories Graph (WCG) which is exploited in this work. Section 4 details the process steps of deriving an "is a" taxonomy from WCG. Section 5 presents the process followed for linking the sub-DAGs resulting from the deriving process. Section 6 describes the assessment performed using the semantic similarity measures to estimate the degree of semantic similarity between two words. The final section is devoted to presenting our conclusions and recommendations for future research.

2. Related work

The extraction of semantic relations, precisely the "*is a*" taxonomic relation, is based on various components of Wikipedia. The literature presents several proposals advanced for this purpose. In fact, relations between concepts are extracted starting from texts, infoboxes, pages, hyperlinks, and Wikipedia categories. This section presents some of the major approaches proposed for taxonomy or ontology construction or derivation. Based on the nature of the Wikipedia components they employed, these approaches can be classified into three broad categories, namely textual, structural, and hybrid approaches. Structural approaches use the hierarchical information of Wikipedia, including WCG, pages graph, and infoboxes. Textual approaches, on the other hand, employ the content of each page and analyze category names. Some hybrid approaches combine aspects from both structural and textual approaches.

2.1. Textual approaches

The page is the most important component in Wikipedia. The majority of pages contain texts that are presented in various semi-structured forms, such as: section, sub-section, and paragraph, and include hyperlinks towards other pages. DBpedia (Auer et al., 2007) is a project that aims to extract structured information starting from Wikipedia pages and binding them to Web data.

YAGO (Suchanek et al., 2007) is a project that connects Wikipedia categories to the synsets of WordNet to form an ontology. The extraction process includes rules and heuristics to extract concepts from Wikipedia categories. Those heuristics serve to analyze category names and check whether their heads are in plural form to consider them as concepts. In order to evaluate the YAGO's performance, the concepts are randomly selected from the extracted ontology. Human judgments are then used to evaluate the correctly extracted relations.

Garcia et al. (2012) proposed an approach using a set of 20 characteristics to identify the "*is a*" relations between Wikipedia categories. They used syntactic characteristics based on the lexical headings of category names. Among those characteristics, they quote "*positionOfHeadFeature*" which returns values between $[-1, 2]$ to categories c_1 and c_2. Those values represent the position of the name head of c_2 in the name of the category c_1. If the head of c_2 is at the end of c_1, then the value is 2; if it is in the middle position, then the value is 1; and if it is at the beginning, then the value is 0. For example, c_1="*French Revolution*" and c_2="*Revolution*"; thus, the value is 2 and the relation between them is qualified as "*is a*". In addition, they proposed features applied to page contents. For example, "*definitionSentenceFeature*" is the characteristic used to recognize the "*is a*" relations from the first sentence. In fact, the first sentence of a page has a special meaning according to the taxonomic applications because it generally contains a definition of the designated concept.

Recently, Hoffart et al. (2013) proposed YAGO2[7] as an extension to YAGO. This system is able to extract the subsumption relations from Wikipedia with a high degree of accuracy by determining whether the words in category heads or articles are in the plural or singular form (e.g. if the name of the category "*Plants*" is in the plural, then the relationships with its descendants are considered as relations of subsumption).

2.2. Structural approaches

Several studies have investigated the automatic knowledge extraction from Wikipedia and, more precisely, categories, infoboxes, and categories graphs. The taxonomic relations generated from Wikipedia have often been evaluated using a set of randomly selected samples and their comparison with other knowledge bases, such as WordNet and Cyc.

Sumida and Torisawa (2008) proposed an approach to extract hyponymous relationships from the structure of a Wikipedia page. They define hyponymy as a hierarchical relationship between a more general term (hypernym) and its more specific instances (hyponyms). For example, "*plant*" is a hypernym of "*flower*" which is its hyponym. The automatic extraction of this relation is carried out for two languages: Japanese and English. The authors used the hierarchy of the titles and subtitles of sections included into the Wikipedia pages.

In the source code of the pages, this hierarchy is easily readable owing to the use of markups, such as "=*" and "=", to mark the title of a subject in Mediawiki[8]. The same system uses "*" to specify the subject of a list of elements. These dependences of section and sub-section were used to automate the extraction of hyponym relationships. This method collects several relations using the following three steps:

- The first step consists in extracting relations using the section titles. The title is coupled to all its direct subordinates to create candidate relations such as "*Louvre/Geography*", "*Louvre/Toponymy*". These extracted relations can obviously include erroneous ones.
- The second step involves the selection of hyponymy relations per application of simple patterns (containing the terms *Type* or *Popular*) to the candidate relations obtained at the preceding step. For example, "*Zoo*" is a Wikipedia page that includes "*Popular animals*", "*Magellanic Penguin*", "*Lion*", etc. When applying step 1, the candidates of the hyponym relation of "*Popular Animals/Magellanic Penguin*" are extracted. The pattern "*Popular*" is applied to the hypernyms candidate and then removed to have results such as "*Animals/Magellanic Penguin*".
- The last step applies a training machine to the relations resulting from the previous step. For each candidate relation, a morphological analysis is applied. Then, the characteristics of each hypernym and hyponym candidate are obtained and placed in a vector that will be given as input to the classifier.

The authors evaluated their work using the Japanese Wikipedia version, which contained 820.074 pages, the classifier TinySVM[9], and the analyzer MeCab[10] with a random selection of 200 hypernym relations to calculate the precision of the relations extracted from the hierarchy. The values of the precision were evaluated and compared to 400 samples selected randomly among the candidate relations.

WikiTaxonomy is a taxonomy extracted from the WCG suggested by Ponzetto and Strube (2007). The categories in this taxonomy are considered as concepts, and the relations between concepts are labeled as "*is a*" or "*non is a*" based on the network connectivity and syntax. This method is an extension of their previous work (Strube and Ponzetto, 2006) where they consider that the categories graph

[7] www.mpi-inf.mpg.de/yago-naga/yago2/.

[8] https://www.mediawiki.org/wiki/MediaWiki.

[9] http://chasen.org/ taku/software/TinySVM/index.html.

[10] http://mecab.sourceforge.net.