Expanding basic ontology from Wikipedia

J.L. Castro, P. Rodríguez and J.M Zurita

University of Granada, Spain

castro@decsai.ugr.es

Abstract. Ontologies are being revealed in recent years as an important resource for the representation and management of knowledge. The construction of an ontology is a slow and complicated process, since it depends on the domain in question and often a wide intervention of the human expert is necessary. In this article we give a basic algorithm to learn a domain ontology in a semiautomatic way, starting from a simple initial ontology. For this, we will expand this basic ontology using an available resource such as Wikipedia. This expansion will be done by expanding the hierarchy of classes and initial objects, constructing a larger hierarchy of classes and objects that better represents the knowledge of the domain that is intended to be learned.

Keywords: Ontology, Learning ontologies, Wikipedia

1 Introduction and Related Work

The new information society in which we are currently immersed requires new technologies that facilitate, above all, the processing of the enormous amount of information that is handled. The rapid growth of the Web, and all the information associated with it, makes necessary new forms and models of representation of all this knowledge in a way that can be stored and managed in an automated way to obtain results of the same.

Ontologies have been developed in Artificial Intelligence since the 1990s, with the main functionalities associated to the sharing and reuse of knowledge. There are important advances in the fields of knowledge engineering, natural-language processing and knowledge representation. They also have an important acceptance in the industry, where they are advancing in the fields of intelligent information integration, cooperative information systems, information retrieval, electronic commerce and knowledge management.

The main structure of an ontology consists of a set of concepts that are related through the relations of hyperonymy and hyponymy, thus constituting a taxonomy of concepts. For example, the concept Country is a broader concept than encompasses the concept of the United States and the concept of Spain. Therefore, we can say that Country is a hyperonymy of the United States and Spain, and that in turn, the latter are hyponymy of Country.

The semantic character of an ontology is enriched by the use of properties on concepts and new relationships between them, which makes possible the intelligent use of the knowledge that stores. Thanks to this type of representation, new knowledge can automatically be inferred. The use of ontologies is especially significant when the knowledge is coming from a specific domain, not a general knowledge. We can therefore call these ontologies specific domain ontologies.

The construction of an ontology is an arduous and expensive process, since the amount of information to be considered, even if we speak of a particular domain, is enormous. In addition, the intervention of an expert in the domain is necessary. The resource "human expert" is expensive and therefore we tend to reduce as much as possible its participation in the construction of the ontology.

There are multiple works that pursue the automatic or semiautomatic construction of an ontology, many of them depart directly from free text, usually text that is found on the web as a source of knowledge. As [5] and [8] who use lexical templates to obtain relationships of hyperonymy and hyponymy through queries made on the web. These methods have the disadvantage that the number of lexical templates can be very high.

Other methods start from the text and directly locate terms and their relations of hyperonymy and hyponymy through the use of clustering, [6]. These clustering methods also use techniques based on text mining and Natural Language Processing (NLP). Other ontology learning methods can be found in [1], [2], [3], [4], [7], [9].

In the present paper we will propose a method for learning concepts and relations of hyperonomy and hyponymy between them, starting from a simple initial taxonomic structure, as a seed. This study aims to improve the difficulties that are presented in the methods based on lexical and clustering templates, such as the excessive number of templates to be considered, or the ambiguity of certain terms during the clustering process; which has repercussions on the completeness and quality of the learning carried out. The proposed method will make use as the main element of a very widespread and clearly automated resource, such as Wikipedia.

In the following section we will show the basic ideas and algorithms on which the proposed method is based. Below we will present a discussion on the main aspects to consider in order to improve and develop this approach in depth.

2 Key ideas and algorithms for expanding a basic ontology

For our purpose, a basic ontology would be composed of classes, subclasses and objects, in a hierarchical structure. Each class would have associated tags and each object as well.

To illustrate the main ideas of the method, we will use throughout the document the basic initial ontology that appears in Fig. 1.

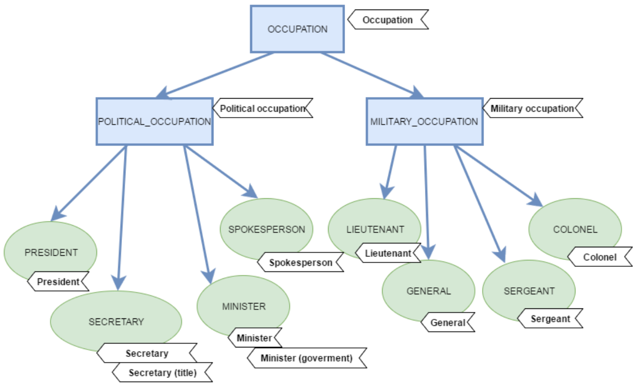

Fig. 1. Example of initial ontology

The Wikipedia has a structure that allows us to establish certain equivalences with the taxonomy of an ontology. See Table 1.

Table 1. Equivalence of structure.

Ontology	Wikipidia
Class	Category
Object	Article/Category
Label	Title

2.1 Preparing and expanding the initial ontology

To begin the expansion, we first need to prepare the initial ontology by dividing it into clusters of siblings (objects) and parent (class), duplicating, if necessary, objects that have more than one parent.

The expansion consists of:

1. Search for objects as Wikipedia articles and addition as classes the categories under which those objects are found.
2. Addition to ontology as new objects Wikipedia articles that are classified as the objects sought.
3. Addition to the ontology of the categories that relate the category of the new objects in Wikipedia with the initial class of the ontology.

Expansion of classes. For each group of brothers, we look for the common categories of lower level found in Wikipedia. To do this, we look for the objects in Wikipedia by their labels. If an object does not appear, we discard it from the expansion leaving it where it was. We choose the common class that best group objects. For this we look for the one that has less number of intermediate categories added to reach the common class. We add to the ontology the intermediate categories obtained from the initial objects. See Fig. 2 and Fig. 3.

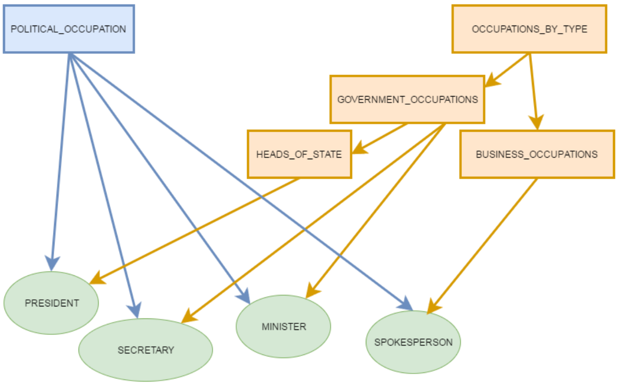

Fig. 2**.** Less common class for objects whose parent is Political_Occupation.

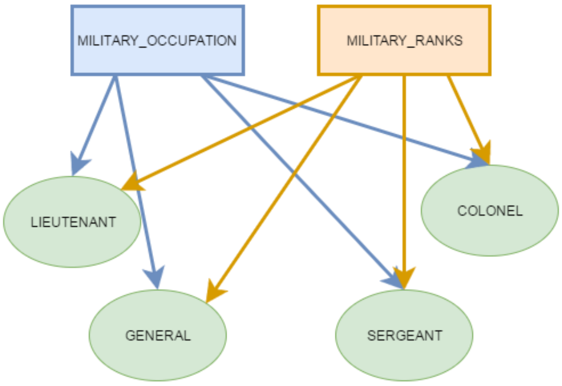

Fig. 3**.** Less common class for objects whose parent is Military_Occupation.

For each initial object found in Wikipedia, we get all its titles and add the titles as labels to their corresponding objects. See Fig. 4.

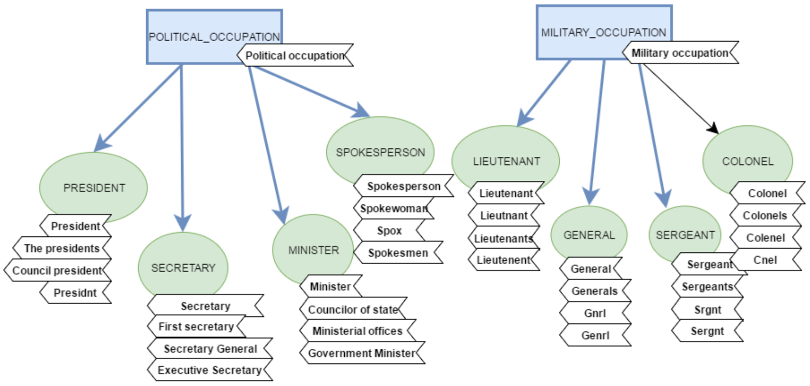

Fig. 4. Expansion of labels for the initial objects.

Expansion of objects. We already have the objects grouped under what Wikipedia considers its common class. Now we are going to add to the ontology new objects that are the same common class. We obtain all the articles of Wikipedia that belong to the first category of the path of categories of each object towards the common class. That is, of the new class structure arising from the previous expansion, we select the first category of each initial object to expand from there. See Fig. 5 and Fig. 6.

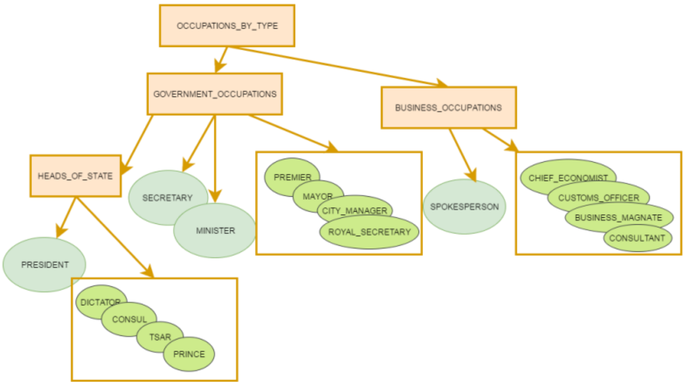

Fig. 5. Expansion of news objects in Political_Occupation.

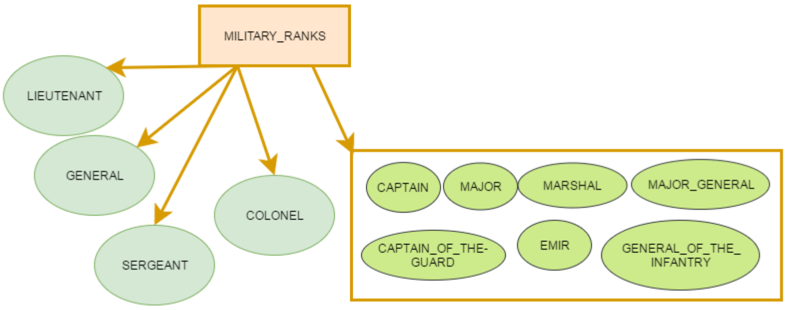

Fig. 6. Expansion of new objects in Military-Occupation.

We look for pages and subcategories under that category and we decide which group of items to select, whether subcategories or pages (articles). To do this, we use the initial objects with the expanded tags and count how many different occurrences of these (of their labels) we have both subcategories and pages.

Connecting category found in Wikipedia with initial class in ontology. After the expansion of objects, for each group of brothers we have a tree of categories up to the common category in Wikipedia. It expands the father of each group of brothers towards the common category of these, found in Wikipedia. The resulting new categories (if found) are added as classes. And the new categories that connect objects with the common category are added as classes in the ontology, preserving the hierarchy. See Fig. 7.

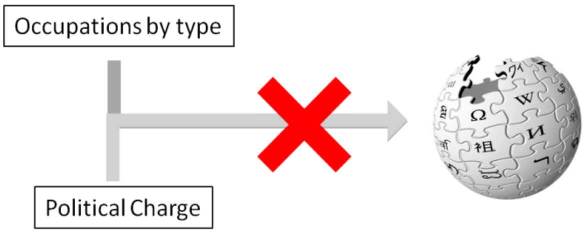

Fig. 7. Connecting the category of Wikipedia with the initial class of the ontology.

The search is divided into two phases: top search and bottom search. There are 3 possibilities: The father is not found on Wikipedia as a category, no connection is found between parent and common category, and the last, there is a path between the father and the common category. If the parent is not found on Wikipedia or there is no connection then we add the expansion of the brothers directly under the father; the common category would remain as a subclass of the parent of its objects. See Fig. 8.

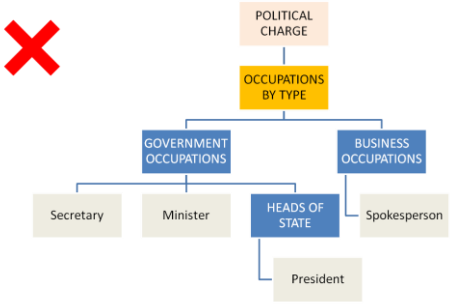

Fig. 8. No connection is found.

If a connection is found between the parent and the common category then we check whether it has been through a higher or lower expansion.

In the first case, the common category must be above the father. We add the categories of the path in the same order that they are in Wikipedia, the parent is inserted at the end of the list. See Fig. 9.

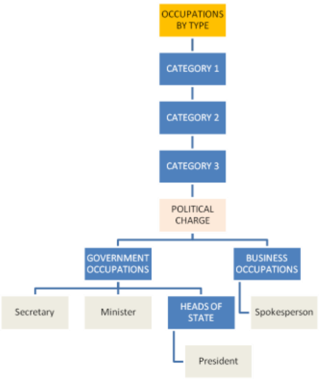

Fig. 9. Higher expansion.

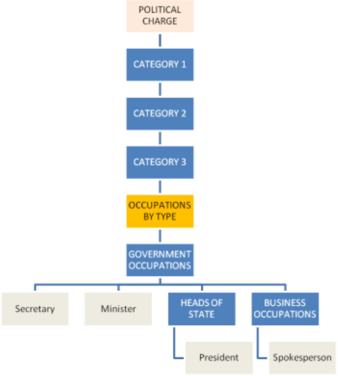

Fig. 10. Lower expansion.

In the second case, the common category should be below the parent. See Fig. 10.

3 Conclusions and future works

In this article we present some ideas and algorithms for the expansion of an initial ontology from a resource such as Wikipedia. The proposed expansion process has the advantage of being automated, with little intervention by the human expert and the disambiguation of concepts is also improved. It has made use of Wikipedia for its structure similar to the taxonomic structure of an ontology.

As future works more immediate to improve these algorithms we propose the followings: There is a particular case that is repeated in numerous occasions when we look for in Wikipedia and is that sometimes, the Wikipedia encompasses a page (article) under a category with the same name. The algorithm must be adapted to work in this situation.

Also, after expanding all the objects of a class, we can find that some object initially grouped with his brothers, after finding it in the Wikipedia, has been isolated under categories very distant of its initial brothers. When we expand the solitary object, the new objects generated may not keep too much relation the their initial sibling objects which can lead to the introduction of “noise” in the expansion. [3], [7].

Acknowledgements This research was supported by the project TIN-2013-4839-R.

References

1. Chernyak, E, Mirkin, B.: A method for refining a taxonomy by using annotated suffix tress and Wikipedia resources. Procedia Computer Science, 31, 193–200 (2014).
2. Kang, Y.B., Haghigh, P.D., Burstein, F.: Taxofinder: A graph-based approach for taxonomy learning. IEEE Transactions on Knowledge and Data Engineering, 28(2):524-536 (2016).
3. Maedche, A., Staab, S.: Ontology learning for the semantic web. IEEE Intelligent Systems, 16, 72-79 (2001).
4. Meijer, K., Frasincar, F., Hogenboom, F.: Ontolearn reloaded: A graph-based algorithm for taxonomy induction. Association for Computational Linguistics, 62, 78-93 (2014).
5. Ortega-Mendoza, L., Villaseor-Pieda, L., Gomez, M.M.: Using lexical patterns for extracting hyponyms from the web. In: MICAI 2007: Advances in Artificial Intelligence, 904-911 (2007).
6. Pantel, P., Lin, D.: Discovering word senses from text. In: 8th ACM SIGKDD: International Conference on Knowledge Discovery and Data Mining, 613-619 (2002).

7. Ros-Alvarado, A.B., Lopez-Arevalo, I., Sosa, V.J.: Learning concept hierarchies from textual resources for ontologies construction. Expert Systems with Applications, 40, 5907-5915 (2013)
8. Sang, E. Extracting hypernym pairs from the web. In 45th Annual Meeting of the Association for Computational Linguistics. (2007).
9. Velardi, P., Faralli, S., Navigli, R.: Ontolearn reloaded: A graph-based algorithm for taxonomy induction. Association for Computational Linguistics, 665-707 (2013).