A Semi-Automatic Method for Domain Ontology Extraction from Portuguese Language Wikipedia's Categories

Clarissa Castellã Xavier[1], Vera Lucia Strube de Lima[1]

[1] Faculdade de Informática – PUCRS, Av. Ipiranga, 6681 – Prédio 32, Porto Alegre, Brazil
{clarissa.xavier, vera.strube}@pucrs.br

Abstract. The increasing need for ontologies and the difficulties of manual construction give place to initiatives proposing methods for automatic and semi-automatic ontology learning. In this work we present a semi-automatic method for domain ontologies extraction from Wikipedia's categories. In order to validate the method, we have conducted a case study in which we implemented a prototype generating a Tourism ontology. The results are evaluated against a manually built Golden Standard reporting 79.51% Precision and 91.95% Recall, comparable to those found in the literature for other languages.

Keywords: ontologies, Wikipedia, semi-automatic ontology extraction.

1 Introduction

According to Wikipedia's project maintainer, Wikipedia is a free multilingual online encyclopedia, non-profit, collaborative, created on January 15, 2001. In April 2010, Wikipedia had more than 555,000 articles in Portuguese language and more than 3,235,000 articles in English.

Wikipedia's documents are organized in a hierarchy of categories that can be understood as a structure of terms, not strictly a tree structure, but a richer representation. This structure allows multiple simultaneous categorization of topics. Some categories may have more than one super-category [1], forming a graph that represents a conceptual network with unspecified semantic relations [2].

In Computer Science ontologies are understood as "an explicit specification of a conceptualization"[3]. There are even simpler definitions for ontologies, such as the W3C[1] consortium featuring ontology as "the definition of terms used to describe and represent an area of knowledge", as well as more complex definitions, such as proposed by Guarino [4] and Smith and Welty [5], who consider classes, properties, instances, axioms and logic to build ontological structures[2].

1 *http://www.w3.org/TR/2003/ WD-webont-req-20030203*

2 We will use the terms "ontology" and "ontological structure" interchangeably, and we will adopt, to ontology, an open approach.

Krötzsch et al. [6] introduce the concept of class in ontologies with "Classes can be compared to Wikipedia's categories: they describe collections of objects and can be organized in a hierarchy". For instance, actor is a subclass of person. As in Wikipedia, the multiple inheritance and even cycles are allowed in the class hierarchy.

Building ontologies is a costly, tedious and error-prone process [7], and the number of domain ontologies currently available remains extremely small [8], scenario that is even worse in Portuguese language [9]. One alternative option for ontologies extraction is to use Wikipedia as data source.

In this context, we present a method for extracting domain ontologies in Portuguese language from Wikipedia's category structure. In order to validate this method we have developed a case study. We have built a prototype that extracts a Tourism ontology containing classes, instances and relations (is-a, located-in) from the Wikipedia database in Portuguese. The results were evaluated by comparing the obtained ontology with a Golden Standard manually built from Wikipedia's Tourism category. These results (79.51% Precision and 91.95% Recall) were promising, demonstrating the feasibility of the proposed method.

The article is organized as follows: Section 2 reviews related work on ontology extraction from Wikipedia. In Section 3 we present our method for domain ontology extraction. The case study, its evaluation and results are presented in Section 4. We end the article with a brief conclusion.

2 Related Work

Several studies, as [1, 2, 8, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19] support our research on ontologies extraction from Wikipedia, especially regarding the feasibility of this task. In this section, we present a brief review of those which provided relevant ideas for our method.

YAGO: A Large Ontology from Wikipedia and WordNet [16] presents an ontology derived from Wikipedia and WordNet. The ontology is populated by facts derived from Wikipedia's category system and infoboxes[3], combined with taxonomic relationships from WordNet. This is made in two stages: first, different heuristics are applied to Wikipedia's data in order to extract facts and candidate entities, and the connection between Wikipedia and WordNet is established. The second stage is the application of quality control techniques. The results were evaluated by human judges, who found that 74 heuristics had precision greater than 95%.

Ponzetto and Strube, in [13], describe an experiment for automatic creation of a taxonomy from Wikipedia's category system. Further articles by Ponzetto and Strube [15]; Zirn, Nastase and Strube [18]; and Nastase and Strube [19] describe methods to automatically distinguish between classes and instances from the taxonomy generated in [13].

Wikipedia's category structure seems to be an excellent source of data for ontologies extraction, since it contains relationships between concepts. In the work of Strube and Ponzetto [12, 13] categories are used to describe concepts. [18, 19] report

3 Infobox is a Wikipedia standard model with basic information about the entity described in the article.

more detailed analysis of category titles, describing the extraction of classes and instances. [13] describes very clearly the steps for identifying is-a and not-is-a relations in the category structure, however, it is not clear how the lexical-syntactic patterns were used.

Data extraction from infoboxes and WordNet seems to be promising in English language. However, currently, this feature does not appear to be useful in Portuguese for two reasons: infoboxes are little used in Wikipedia's Portuguese version and the available versions of WordNet in Portuguese language does not have a significant amount of content.

The reviewed papers reported experiments based on the full content of the encyclopedia and not on a specific category. Furthermore, the reviewed studies performed the extraction of content in English only. The performance of a work that extracts an ontology from the Portuguese version of Wikipedia should take in consideration the characteristics of this release, such as size, less use of models and the different spellings of the same term, given the inherent differences between Brazilian, African, Asian and European Portuguese.

3 Extraction Method

This section presents our semi-automatic method for extracting domain ontologies from Portuguese Wikipedia's category system.

The input to the extraction is Wikipedia's database, particularly the tables with information about categories. After applying the method, we obtain a domain ontology described in OWL, containing classes, instances and relations. In the following subsections we describe the stages of the semi-automatic extraction method illustrated in Figure 1.

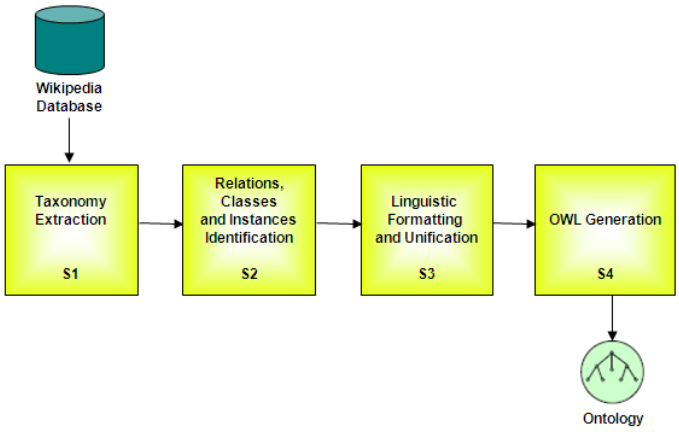

Fig. 1. The method input is Wikipedia's database. It generates a domain ontology containing classes, instances and relations.

3.1 Stage 1 - Taxonomy Extraction

The goal is to obtain a taxonomy where concepts are category titles and the hierarchical relation between concepts is established by the category structure organization in the database. The input here is Wikipedia's database (particularly the tables containing data related to the category structure), the category used as source for extraction and the depth of the query. The output is a taxonomy with the category structure.

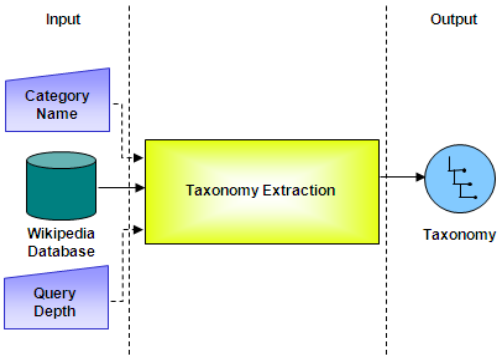

Fig. 2. First Stage (S1) generates a taxonomy where the hierarchical relation is established by the category structure in Wikipedia.

Wikipedia covers different domains of knowledge and the organization of its categories enables the connection of concepts that belong to more than one domain. Since our goal is to obtain a domain ontology, we must limit the selection of subcategories to a limited depth, in order to obtain the largest number of concepts but also ensuring that the selected categories belong to the original domain.

The result is the selection of the subcategories of the chosen category. We emphasize that the number of levels to be searched in the category tree depends on the characteristics of the category to be used as the basis for extraction. Therefore, this number must be previously defined by the ontology engineer.

3.2 Stage 2 - Relations, Classes and Instances Identification

This stage (S2) performs the extraction of relations, new classes and instances, through the analysis of the concepts in the taxonomy generated in the previous step (see Figure 3).

An ontology composed by classes, instances and relations results from this stage. To reach this structure it is necessary previously: to define the relations to be extracted; to define the heuristics to be used to extract relations and to distinguish between classes and instances.

In this stage, each concept is analyzed, checking if there is, embedded in it, a semantic relation other than hyponymy. In this case, the is-a relation is replaced with a new one; and we check if the class name remains the same, or if it should be replaced by a new class and instance.

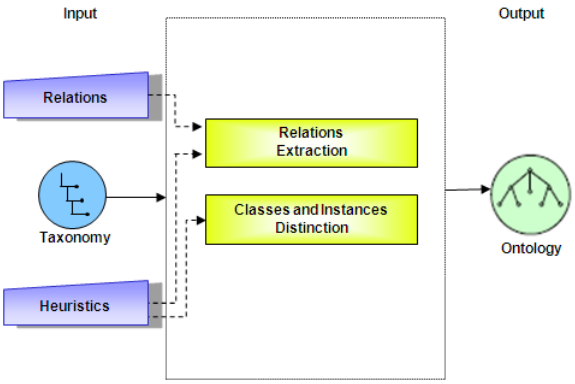

Fig. 3. Second Stage (S2) extracts relations, new classes and instances from the taxonomy generated in S1.

The activities in this stage were based on the Category Heuristics reported in [16] and on methods described in [15, 18, 19] to automatically distinguish between classes and instances in a taxonomy.

3.2.1. Relations Extraction

According to Gruber [20], relations are a set of tuples that represent a relationship between objects in a universe of discourse.

Strube and Ponzetto claim that Wikipedia's category structure does not constitute a taxonomy with a well-formed subsumption hierarchy, but a thematically organized thesaurus [13]. The exclusive use of hyponymy relations (is-a) does not reflect the semantic relationship between concepts in the categories taxonomy extracted from Wikipedia. The use of other relations in conjunction with is-a is essential to accurately describe the semantic connections between concepts. For example, the category of “Capital in Asia” is registered under Capital (is-a), but the category Philosophy is under the category “Abstraction and Belief” (deals-with) and also under “Humanities” (is-a) [13].

The choice for which relations should be extracted is directly related to the domain represented in the ontology. It is first necessary to examine the taxonomy obtained in the previous step to then define, from its concepts, which relations represent the semantic relationships between classes. For example, in the Tourism taxonomy, some categories have location relations embedded in their title, such as "Zoos in Germany", which can be represented by the relationship "Zoos located-in Germany".

3.2.2. Distinction between classes and instances

Instances represent the objects of the domain on which is our interest, while classes are interpreted as sets of instances [21]. To characterize an instance, we suggest the use of these characteristics [22]: instances are proper nouns, which means that they should be capitalized and instances are unique, which implies that they should not have hyponyms; it is meaningless to have an instance of an instance.

At this point, the concepts selected in the previous stage, as having a different relationship than hyponymy, are analyzed in order to check if it is necessary to create new classes and instances. For example, we have identified in th category “Zoos in

Germany" the relationship "Zoos located-in Germany". From this point, we characterize "Zoos" as a new class and "Germany" as an instance.

The heuristics used to analyze the category title should be manually defined in advance, because they depend on: domain, relations to be extracted and the category structure used as source.

3.3 Stage 3 - Formatting and Linguistic Unification

In this stage (S3) the ontology names are standardized. As a first task we format titles, so they can be represented in OWL. This task is performed in three steps:

1. Remove special characters;
2. Replace spaces with underscore;
3. Convert all characters to lowercase.

The second task unifies different spellings, considering that Wikipedia's Portuguese version, according to its manual of style[4], does not use a specific version of the common language, regardless their country of origin. Thus, the same term may be registered more than once with different spellings, with the differences inherent to all variants of Portuguese.

3.4 Stage 4 – OWL Generation

The last stage (S4) generates the OWL description of the ontology obtained in previous stages. The OWL description allows the visualization and refinement of the extracted ontology in ontology editors such as Protégé[5], as well as easy access by other applications.

4 Case Study

In this section we describe a case study developed to validate the method presented in the previous section. We have conducted an experiment extracting a Tourism ontology containing classes, instances and relations (is-a, located-in) from Portuguese Wikipedia's Tourism category[6]. The generated ontology was confronted against a Golden Standard and the metrics Precision and Recall were calculated.

4.1 Prototype

The prototype developed to conduct the case study was implemented in PHP,

4 *http://pt.wikipedia.org/wiki/Wikipedia:Livro_de_estilo*
5 *http://protege.stanford.edu/*
6 *http://pt.wikipedia.org/wiki/Categoria:Turismo*

accessing Portuguese Wikipedia's database in Mysql[7] and generating an OWL file. The prototype was designed in four steps, according to the method being validated. Next we describe the prototype steps.

Step 1 – Categories Selection. In order to select the depth of the Tourism category query, we have manually performed an analysis of its subcategories graph and from this observation, we decided to set the search in three levels of subcategories, trying to get as many concepts as possible without exceeding the Tourism domain.

Upon completion of this step, we get a taxonomy where the concepts are the categories titles and the hierarchical relationship is established by how the category structure was organized in Wikipedia's database.

Step 2 – Relations, Classes and Instances Identification. To accomplish this task we have previously defined the relations and instances to be extracted and the heuristics to be used. We have noticed that many of these semantic relations would be better represented by the located-in relation. For example, the two categories with the largest number of subcategories are "Transportation by country"[8] and "Tourism by country"[9], which are categories whose semantic content is related to location. In addition, some categories present located-in relationship embedded in their title, for example, "Spas in Brazil"[10] which in our view includes the relationship "Spa" located-in "Brazil".

From this definition, we found that located-in relations did not occur only among classes, but between class and instance of a place. In the relation "Spa located-in Brazil", we classify "Spa" as class and "Brazil" as instance. To accomplish the task of identifying location relationships and to make the distinction between classes and instances we have proposed four heuristics:

Heuristic 1: infers the existence of located-in relationships in the subcategories of categories whose title contains the words "country", "city", "province" or "state". For example, "Touristic attractions in Curitiba"[11] is subclass of "Tourist attractions by city"[12]. The application of the rule generates the instance "Curitiba" and the relation "Touristic attractions by city" located-in "Curitiba" and, also erases the class "Touristic attractions in Curitiba".

Heuristic 2: infers located-in relationships in categories containing certain prepositions[13] in its title. For example, from "Airports in Argentina"[14] the application of this rule generates the class "Airports", the instance "Argentina", the relation "Airports" located-in "Argentina" and erases the class "Airports in Argentina".

Heuristic 3: for classes containing only one lexical item in their title, search in Wikipedia's database for connections between the correspondent category and other

---

7 Database dump obtained in http://download.wikimedia.org/backup-index.html from Wikipedia's Portuguese version of January 05, 2009.

8 "Transporte por País" in Portuguese.

9 "Turismo por País" in Portuguese.

10 "Termas do Brasil" in Portuguese.

11 "Atrações turísticas de Curitiba" in Portuguese.

12 "Atrações turísticas por cidade" in Portuguese.

13 In Portuguese: "de/do/da" and "em/no/na".

14 "Aeroportos da Argentina" in Portuguese.

categories related with locations. If the connection is found, the class is transformed in instance and a located-in relation is established with its super class. I.e, “Krakow”[15] is subclass of “UNESCO World Heritage”[16] and "Cities of Poland"[17]: this creates the relation “UNESCO World Heritage” located-in “Krakow”.

Heuristic 4: performs a quality control by excluding wrong mappings. If an instance was also mapped as a class, the mapping as instance is eliminated.

Step 3 – Spelling Unification. At this stage, we standardize classes and instances names, allowing the OWL creation in the next step.

To do so, we perform a function that replaces string "cç" (Portugal’s Portuguese) spelling for "ç"; we remove words accents, ie. replacing "ã" with "a"; we convert uppercase to lowercase; we replace blanks by underscore.

Step 4 – OWL Generation. Finally, the last stage generates an OWL file containing the extracted ontology description.

The ontology created with the prototype execution consists of 165 classes and 156 instances.

4.2 Evaluation

In order to evaluate the results obtained with the case study reported in the previous session, we compute Precision and Recall. From these, we investigate some of the causes of successes and mistakes of the prototype, examining the similarities and differences between the Golden Standard and the ontology being evaluated.

The Golden Standard used to calculate the metrics was manually constructed from the Tourism Category Network, revised and refined by a linguist. The Golden Standard was developed following three steps:

1. Tourism Category Structure export in three levels into a taxonomy in OWL.
2. Manual construction of the ontology, from the taxonomy generated in the previous step.
3. Review and refinement of the ontology by a linguist.

Confronting both ontologies as a whole, we have obtained 79.51% Precision and 91.95% Recall. These results are satisfactory and comparable to those found in the literature for other languages.

The main differences between the ontology generated by the prototype and the Golden Standard are in the is-a relation mapping. We have achieved for this analysis 73.03% Precision and 91.98% Recall. The main cause of differences was produced by the Heuristic 2.

For example, in the Golden Standard the class “hotels”[18] is a subclass of “lodging_facilities”[19], while in the extracted ontology the class “hotels” is also a

15 “Cracóvia” in Portuguese.
16 “Patrimônio Mundial da UNESCO” in Portuguese.
17 “Cidades da Polônia” in Portuguese.
18 “hotéis” in Portuguese.
19 “meios_de_hospedagem” in Portuguese.

subclass of “tourism_in_south_america”[20]. This happened because applying the Heuristic 2 in the class “hotels_from_brazil”[21]: generates the class "hotels"; generates the instance "Brazil"; generates the relation “hotels” located-in “Brazil”; places the class “hotels” as subclass of “tourism_in_south_america” (original position of the class “hotels_from_brazil”).

The best results were achieved by the instance mapping, obtaining 99.35% Precision and 94.44% Recall. Only one instance was wrongly mapped by prototype: “superhighway”[22]. Heuristic 3 caused this failure, since “superhighway” contains only one word in its title and is connected to the “Rio de Janeiro City Transports”[23].

5 Conclusion

We have presented a method for semi-automatic extraction of domain ontologies from Portuguese Wikipedia. Toward this end we exploited Wikipedia’s category structure and the categories names as source for ontology components extraction. The method input is Wikipedia’s database, particularly the tables containing information about categories. After applying its four stages, it generates a domain ontology described in OWL.

The method was validated through a case study that produced a Tourism ontology. In order to evaluate this ontology, a Golden Standard was manually constructed from the Tourism Category structure, revised and refined by a linguist. The results were satisfactory: 79.51% Precision and 91.95% Recall.

This evaluation showed that the main differences between the ontology generated by the prototype and the reference ontology are in the mapping of is-a relations. One of the reasons that led to this inequality was the exclusion of repeated classes in the Golden Standard by the human revisor. The prototype had its best performance in the instance mapping.

From this experience, we point that the appropriate definition of heuristics is the key point for implementing the method successfully.

We believe that the use of a Golden Standard is suitable and widely used in this type of problem. However, we are working to find other alternatives for evaluation. For example, it would be interesting to perform extrinsic evaluation using information retrieval systems.

Future work includes the method improvement, seeking its automation. To accomplish this, we point to the automation of the search depth obtention in the method first stage and the use of other components of the encyclopedia as text and links between articles as data source.

---

20 “turismo_na_america_do_sul” in Portuguese.

21 “hoteis_do_brasil” in Portuguese.

22 “supervia” in Portuguese.

23 “Transportes da cidade do Rio de Janeiro” in Portuguese.

References

1. Zareen, S. et al.: Wikipedia as an Ontology for Describing Documents, In: Proceedings of the Second International Conference on Weblogs and Social Media (2008).
2. Wu, F., Weld, D. S.: Autonomously semantifying wikipedia. In: Proceedings of the Sixteenth ACM Conference on Conference on information and Knowledge Management (CIKM '07), pp. 41--50. ACM, New York, NY (2007)
3. Gruber, T. R: A translation approach to portable ontology specifications. In: Knowl. Acquis. 5, 2, pp. 199--220 (1993)
4. Guarino, N.: Formal Ontology. In: Information Systems: Proceedings of the 1st International Conference, 1st. IOS Press, Trento, Italy (1998)
5. Smith, B.; Welty, C.: FOIS introduction: Ontology- towards a new synthesis. In: Proceedings of the international Conference on Formal ontology in information Systems - Volume 2001 (Ogunquit, Maine, USA, October 17 - 19, 2001). ACM, New York, NY (2001)
6. Krötzsch, M., Ci´c, D., Völkel, M.: Wikipedia and semantic web - the missing links. (2005).
7. Maedche, A. D.: Ontology Learning for the Semantic Web. Kluwer Academic Publishers (2002)
8. Hepp, M., Bachlechner, D., Siorpaer, K.: Harvesting Wiki Consensus - Using Wikipedia Entries as Ontology Elements. IEEE Internet Computing v. 11-5, pp. 54--65 (2007)
9. Lima, V., Nunes, M., Vieira, R.: Desafios do Processamento de Línguas Naturais. In: SEMISH - XXXIV Seminário Integrado de Software e Hardware. Anais do XXVII Congresso da SBC, Rio de Janeiro (2007)
10. Völkel, M., Krötzsch, M., Vrandecic, D., Haller, H., Studer, R.: Semantic Wikipédia. In: Proceedings of the 15th international Conference on World Wide Web, pp. 585--594 (2006)
11. Wu, F., Weld, D. S.: Automatically refining the wikipedia infobox ontology. In: Proc. of the 17th Int. Conf. on WWW, pp. 635--644. ACM, New York, NY (2008)
12. Ponzetto, S., Strube M.: Knowledge Derived from Wikipedia for Computing Semantic Relatedness. Journal of Artificial Intelligence Research, v. 30, pp. 181--212 (2007)
13. Ponzetto, S. P. and Strube, M.. Deriving a large scale taxonomy from Wikipedia. In Proceedings of the 22nd National Conference on Artificial intelligence - Volume 2, pp. 1440--1445. AAAI Press (2007)
14. Nastase, V; Strube, M.: Decoding Wikipedia Categories for Knowledge Acquisition. In: Twenty-Third AAAI Conference on Artificial Intelligence, pp.1219--1224 (2008)
15. Ponzetto, S. P. and Strube, M.: WikiTaxonomy: A Large Scale Knowledge Resource. In: Proceeding of the 2008 Conference on ECAI 2008, pp. 751--752. M. Ghallab, C. D. Spyropoulos, N. Fakotakis, and N. Avouris, Eds. Frontiers in Artificial Intelligence and Applications, vol. 178. IOS Press, Amsterdam, The Netherlands. (2008)
16. Suchanek, F. M., Kasneci, G., Weikum, G.: YAGO: A Large Ontology from Wikipedia and WordNet. Web Semant. 6, 3, pp. 203--217 (2008)
17. Syed Z.; Finin T.; Joshi A.: Wikipedia as an Ontology for Describing Documents. In: Proceedings of the International Conference on Weblogs and Social Media (2008).
18. Zirn, C., Nastase, V., Strube, M.: Distinguishing between Instances and Classes in the Wikipedia Taxonomy. In: ESWC '08, pp.376--387 (2008)
19. Nastase, V., Strube, M.: Decoding Wikipedia Categories for Knowledge Acquisition. In: AAAI '08, pp. 1219--1224 (2008)
20. Gruber, T. R.: Ontolingua: A Mechanism to Support Portable Ontologies. Technical Report (1992)
21. Horridge, C., Knublauch, H., Rector, A., Stevens, R., Wroe, C.: A practical guide to building OWL Ontologies using the Protégé OWL Plug-in and CO-ODE Tools, http://www.co-ode.org/resources/ tutorials/ProtegeOWLTutorial.pdf (2004)
22. Miller, G. A., Hristea, F.: WordNet nouns: Classes and instances. In: Computational Linguistics, 32--1, pp. 1--3 (2006)