Unsupervised Knowledge Extraction for Taxonomies of Concepts from Wikipedia

Eduard Barbu
Center for Mind/Brain Sciences
Rovereto
Trento, Italy
eduard.barbu@unitn.it

Massimo Poesio
Center for Mind/Brain Sciences
Rovereto
Trento, Italy
massimo.poesio@unitn.it

Abstract

A novel method for unsupervised acquisition of knowledge for taxonomies of concepts from raw Wikipedia text is presented. We assume that the concepts classified under the same node in a taxonomy are described in a comparable way in Wikipedia. The concepts in 6 taxonomies extracted from WordNet are mapped onto Wikipedia pages and the lexico-syntactic patterns describing semantic structures expressing relevant knowledge for the concepts are automatically learnt.

Keywords

wikipedia, unsupervised knowledge acquisition, taxonomy

1 Introduction

A crucial phase in ontology acquisition from text is the extraction of relevant knowledge for ontology concepts, the focus of the current work. Our framework extracts in an unsupervised way knowledge for a set of concepts hierarchically ordered. For example, for the concept **bewick's swan**, one of the concepts in bird taxonomy, some extracted properties are: *have few natural predator*[1], *live in water*, *is a small Holarctic swan*. From a logical/ontological point of view the extracted knowledge can be classified as: quantifier restrictions (e. g. *most birds build nests*), parts of the instances of the concepts in the taxonomy (e.g. *small head* and *long thick mane* for the concept **shetland pony**), alternative classification of the concepts in the taxonomy (*herd animal* and *social creature* for the concept **horse**), etc.

The knowledge relevant for concepts can be automatically extracted from a variety of sources: dictionaries, databases, corpora, web directories and others. Recently, Wikipedia drew the attention of various research groups as a goldmine resource for information retrieval [3], information extraction [9] and ontology building [8].

There are some characteristics that make Wikipedia an appropriate resource for information extraction. Firstly, its coverage is impressive: the English Wikipedia has almost three million articles currently maintained and updated by thousands of voluntary contributors, thus surpassing any other encyclopedia in history. Secondly, the style of writing Wikipedia articles is more homogeneous than the mixed bag of styles one encounters in general corpora or in unrestricted text found on the web. Thirdly, Wikipedia has a large network of links, categories and info-boxes allowing a combination of techniques for information extraction.

This paper introduces a novel method for acquisition of knowledge for taxonomies of concepts from the raw Wikipedia text. We assume that similar concepts (i.e. those classified under the same node in a taxonomy) are described in a comparable way in Wikipedia. More precisely, we suppose that the relevant knowledge of these similar concepts is expressed using equivalent surface patterns. The learning process starts with the generation of concept hierarchies from WordNet. The concepts in each hierarchy are mapped onto Wikipedia pages and the knowledge appropriate to the concepts is automatically extracted at a precision ranging from 55 to 66 percents depending on the taxonomy.

The remaining of the paper is organized as follows. In section 2 we present the mapping of concept taxonomy onto Wikipedia pages and discuss the algorithm for knowledge extraction. Section 3 presents, evaluates and discusses the results. Section 4 compares our work with related approaches and the last section summarizes the results and concludes the paper.

2 Knowledge Extraction for Taxonomies of Concepts

The knowledge extraction precision depends on the accuracy of the classification of Wikipedia pages. Each concept from the taxonomy should be precisely mapped on the corresponding Wikipedia article. Therefore, to generate the taxonomy of concepts and map the generated taxonomy onto Wikipedia articles we follow the next steps:

- First, we pick a concept of interest representing the higher level node of the taxonomy to be extracted and map it onto a WordNet synset. For example, if you have chosen the concept **dog** and you want to get the sense corresponding to the animal, you map the concept to the sense number 1 in WordNet.

[1] In this paper the concepts will be typed in **bold** and the properties in *italics*

International Conference RANLP 2009 - Borovets, Bulgaria, pages 28–32

- Second, the hyponymy (sub)tree having as root the concept chosen in the previous step is produced and the concepts in the tree are mapped onto Wikipedia pages. As others have shown [5] the best mapping heuristic is to choose that member of a synset which has the sense number 1 . Even so, the ambiguity problem is not completely solved. For it is possible that concepts having low or no ambiguity in WordNet to be highly ambiguous in Wikipedia. Fortunately, in this case the Wikipedia server returns a page having a standard structure and allows us to reject the ambiguous concept or to guess the right mapping. The disambiguation is performed concatenating the ambiguous concept with each of its WordNet hyperonyms and searching again in Wikipedia until an unambiguos entry is found. For example, the concept **buckskin** appears in two synsets in WordNet and in 8 possible entries in Wikipedia. Because we are interested in the sense of **buckskin** having the hyperonym horse we concatenate the two words (buckskin_(horse)) and send the new entry to Wikipedia server. Fortunately, in this case no ambiguity results and the correct mapping is automatically performed.

The generated taxonomy is used as input by the system in Figure 1. The Extracted taxonomy is mapped onto the Wikipedia pages (the first part of Figure 1) and the pipeline of the system is made by a set of modules, each of them working on the output produced by the preceding module in the pipeline.

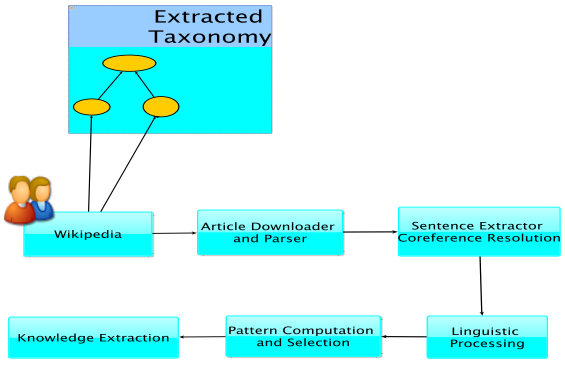

Fig. 1: *The pipeline of the system for knowledge extraction*

The module **Article Downloader and Parser** downloads the Wikipedia articles corresponding to the categories in the taxonomy. From the rough downloaded content of Wikipedia articles we eliminate the useless html tags and the head structure of the article is recovered (e.g. to each higher order head in the article its corresponding text is assigned). In addition, the module eliminates the content of some heads not used by the system, like: Links, Miscellaneous, See also.

The next module, **Sentence Extractor and Co-Reference Resolution**, extracts from the Wikipedia text of an article all sentences containing references to the title concept. The idea behind extracting all sentences containing the title concept is that these sentences express in a direct way relevant information about the categories in the taxonomy. To extend the range of the sentences extracted, the module performs a basic co-reference resolution. It assumes that pronouns like their, it, he, they found within the first three words of a sentence refer back to the title concept. Further, all references at the beginning of a sentence (within the first five words) to any concept in the taxonomic chain of the title concept are also extracted.

Then the module **Linguistic Processing** performs part-of-speech tagging, lemmatization and term identification for the extracted sentences. In order to harvest multi-word expressions and to achieve a better generalization across multiple similar sentences we use the following regular expression of a term definition:

$$(NPrep)?((Adv)?Adj) * (Noun)+$$

The abbreviation NPrep denotes a noun preposition and the straightforward abreviations Adv and Adj denote a adverb an an adjective respectively. The output of this module is a list of sentences in simplified term form (where the terms containing the title concepts are replaced with the generic label TitleConcept and the rest of the terms are replaced with the label T).

The task of the module **Pattern Computation and Selection** is to identify the patterns expressing relevant knowledge for the title concepts. This module has two sub-modules: the first one is called **Pattern Generation** and computes candidate patterns. The second one is named **Pattern Ranking and Selection** and it implements heuristics for ranking and selecting the relevant patterns. The idea behind pattern generation is that the patterns originated should express knowledge characteristic to similar concepts. We judge concepts as similar if they are classified under the same node in the taxonomy and we assume that the relevant knowledge of similar concepts is stated in using the same lexico-syntactic patterns. Therefore, one expects the patterns expressing knowledge of these concepts to appear in the extracted Wikipedia sentences for more than one concept. To produce candidate patterns the Cartesian Product between all sentences in simplified term form (as outputted by the **Linguistic Processing** module) belonging to each pair of similar concepts is performed. For each pair of sentences in the Cartesian product we consider as candidate patterns the longest common substring including the title concept between the sentences. The submodule **Pattern Ranking and Selection** filters the patterns produced by the sub-module **Pattern Generation**. We assume that the best patterns have the shape given by the following regular expression form:

$$(TitleConcept|T)(.+)(T|TitleConcept).$$

Thus we accept the following patterns: "T of TitleConcept be T", "TitleConcept be T", "TitleConcept be design by T" and reject the next patterns: "in

T , TitleConcept be", "of TitleConcept, T". While the former patterns have both topic (what is being talked about; it always contains the TitleConcept) and focus (what is being said about the topic), the latter are incomplete, missing either topic or focus, thus being useless for information extraction. We also reject all patterns having a frequency lower than an experimentally determined threshold.

The module **Knowledge extraction** extracts knowledge for the concepts in taxonomy using the patterns voted in the previous step. For example, applying the voted pattern "TitleConcept consists_of T" to one of the sentences in the entry of the concept **knife** we get part relations:

- **knife** *consist_of a blade*

Moreover, applying the pattern "TitleConcept be_use_in T" to the entries corresponding to the concepts **razor** and **sickle** we extract the function relations:

- **razor** *be_use_in carpentry*
- **sickle** *be_use_in druidic ritual*

3 Results and discussions

3.1 Experimental setup

The input to the knowledge generation experiment is a set of six taxonomies extracted from WordNet as explained in the previous section. The root nodes of taxonomies are three animals (**Horse**, **Dog**, **Bird**), two vehicles (**Aircraft** and **Boat**) and one tool (**Cutlery**). The distribution of concepts for each taxonomy together with examples of concepts is given in Table 1. The number of concepts in the six taxonomies varies from a minimum of 34 concepts to a maximum 128 concepts with an average number of 64 concepts per category. The encyclopedia entries corresponding to the taxonomies categories are downloaded with the software module WWW::Wikipedia. The Wikipedia text is part-of-speech tagged and lemmatized with TreeTagger, a language independent POS tagger.

3.2 Pattern Voting

Table 2 shows examples of patterns voted for each of the six taxonomies. Inspecting the table we observe that a pattern voted in all taxonomies is "TitleConcept be T". This pattern is present in almost all articles in Wikipedia and it is usually found in the first three sentences of the abstract. Included in the term connected with the title concept by the verb to be there is a noun phrase giving the taxonomic classification of the title concept together with other interesting information. However, the taxonomic classification extracted with the help of this pattern is not always found among the superordinate terms in the taxonomy we started with. For example, the extracted superordinate for the concept **red_eyed_vireo** is **songbird**. In WordNet the relevant superordinates of the concept **red_eyed_vireo** are: **oscine**, **passerine** and **bird**, none of which is **songbird**.

Taxonomic Root	Number of Concepts	Examples
Aircraft	34	**monoplane**, **seaplane** **airliner**, **stealth_fighter**
Boat	30	**wherry**, **fireboat** **motorboat**, **steamboat**
Horse	34	**tarpan**, **shetland_pony** **percheron**, **palomino**
Dog	128	**belgian_sheepdog**, **collie** **rottweiler**, **dalmatian**
Bird	121	**crossbill**, **oscine** **nightingale**, **tailorbird**
Cutlery	34	**knife**, **chisel** **sickle**, **razor**

Table 1: *The roots of the extracted taxonomies and concept examples*

As we expected, some of the voted patterns express knowledge specific to the concepts in certain taxonomies. For example, the pattern "T build TitleConcept" is related to concepts in the taxonomy **Aircraft** and the pattern "TitleConcept eat T" is specific to the concepts in the taxonomy **Bird**[2] . In the first case, the knowledge extracted are constructors of aircraft models like: *Pan Am One* or *Edison*. In the second case, the properties obtained are kinds of food (*insects*, *snail*) consumed by different types of birds.

Taxonomic Root	Examples of voted Patterns
Aircraft	TitleConcept be T T use TitleConcept T build TitleConcept
Boat	TitleConcept be T TitleConcept use T TitleConcept have T
Horse	TitleConcept be T TitleConcept be use in T TitleConcept require T
Dog	TitleConcept be T TitleConcept need T TitleConcept also know as T
Bird	TitleConcept be T TitleConcept forage on T TitleConcept eat T
Cutlery	TitleConcept be T TitleConcept consist of T TitleConcept be T with T

Table 2: *Examples of extracted patterns for taxonomy classes*

3.3 Knowledge Evaluation

In the Table 3 we give examples of the generated knowledge for three concepts: **andean_condor**, **air-**

[2] Although we expected that the second pattern "TitleConcept eat T" to appear also in the concepts of the taxonomies **Dog** and **Horse** it turned out that it did not appear or it was not voted as relevant.

ship and **knife** belonging to the taxonomies **Bird**, **Aircraft** and **Cutlery** respectively.

Concept	Examples of Properties
andean_condor	*be_find_in South_America* *be_call the Argentinean_Condor* *Vultur gryphus*
airship	*use dynamic helium volume* *have a natural buoyancy* *be_know_as dirigible*
knife	*consists_of a blade* *come_in many forms* *make_of copper*

Table 3: *Examples of extracted properties for three concepts*

Two raters evaluate the quality of the generated knowledge using a 3-point scale:

- Ideal Knowledge - (2 points). The extracted properties are necessary for the concepts in the taxonomy. They should be part of an ideal list of properties for the taxonomy concepts (e.g. *is omnivorous* for the concept **australian magpie** or *consists of a blade* for the concept **knife**)
- Partially Correct - (1 point) if the extracted properties correctly describe the taxonomy concepts but are not among their ideal list of properties (e.g. *is related to butcher birds* or *described by English Ornithologist John Latham* for the concept **australian magpie**)
- Incorrect Knowledge - (0 points) if the extracted properties do not apply in any way to the category (e.g. the property *number* for the concept **knife** or the property *be on average* for the concept **andean condor**).

The precision of the extracted knowledge is computed using the following formula.

$$Precission = \frac{2N_{IK} + 1N_{PC}}{2N_{Properties}}$$

where

- N_{IK} counts the number of ideal knowledge labels
- N_{PC} represents the number of partially correct labels
- $N_{Properties}$ counts all properties evaluated

Approximately 10 concepts per category are chosen for evaluation. When the two raters disagreed about a label the judge solves the disagreement adding the final label. The inter-rater agreement is computed using the Kappa score [7] and the precision is computed for the judge scores (see table 4).

Each property generated in the rater file was annotated with a type (e.g. classification property, part property, behaviour property, etc.). For the concepts in all taxonomies the algorithm generates part properties (e.g. *leg* for the concept **king vulture**, *blade* for the concept **knife**) and classification properties

Taxonomic Root	Kappa Score	Precission
Aircraft	0.62	0.55
Boat	0.65	0.57
Horse	0.62	0.63
Dog	0.65	0.66
Bird	0.68	0.60
Cutlery	0.79	0.61

Table 4: *The inter-rater agreement and the precision for the extracted knowledge*

(e.g. *medium-large grebe* for the concept **red_necked grebe** or *scent hound* for the concept **beagle**). Then, depending on the taxonomy, the algorithm generates different types of properties. For example, for all animals (the concepts in the taxonomies dominated by **Horse**, **Dog** and **Bird**) a common property type generated is Behaviour (e.g. *sensitive to insecticide* for the concept **greyhound** or *builds a large nest* for the concept **bald eagle**). For tools a common generated property type is the function (e.g. *used by barbers* for **razor** or *used in druidic ritual* for **golden sickle**). Interestingly enough, some extracted knowledge are rules, like: *most birds build nests* or *most helicopters have a single main rotor.*

4 Related Work

With the advent of new information sources many teams are developing methods for large-scale information extraction taking advantage of the huge amounts of unstructured text currently available. In this framework relevant is the work of Pasca ([4]) who exploits both query logs and Web documents to acquire instances and knowledge for open domain classes.

Recently the potential of Wikipedia for information extraction in general and knowledge extraction in particular was acknowledged by many research groups. The methods that use Wikipedia for knowledge extraction can be grouped in two major classes. The first class of methods takes profit of the internal link structure and the structured information in Wikipedia (e.g. infoboxes or templates), while the second class of methods use Wikipedias raw text.

Representative for the second class of methods is the work of [6]. They acquire from Simple English Wikipedia (an Wikipedia variant intended for people whose first language is not English) patterns expressing the semantic relations linking nouns in Princeton WordNet 1.7 (hyperonymy, hyponymy, holonymy and meronymy). Then they gather new instances for these relations improving in this way the WordNet coverage. The reported precision for the newly extracted relationships is between 60 and 70 depending on the relation. A direct comparison between their system and our system is not possible because, in the first place, the framework they use is weakly supervised, while our framework is completely unsupervised. Secondly, their system is tuned to acquire certain kinds of relations (hyperonyms, parts), while our framework does not make any assumption about the relations that should be extracted. However, there is an important

overlap between the patterns for hyperonyms and part relation generated by both methods.

Much sophisticated frameworks for relation acquisition from Wikipedia include the work of [2] who uses a dependency parser to extract hyponymy relations from Wikipedia sentences containing the verb to be. Our approach is different from the other methods mentioned in the way we make use of the Wikipedia text to generate concept knowledge. We do not identify patterns by defining a certain relation using seeds, as it is the standard procedure in CL after the seminal work of Hearst [1]. We assume instead that similar concepts are described in similar ways in encyclopedia-like resources. If the main assumption behind the work of Hearst is that semantic relations can be mapped with a certain precision on lexico-syntactic patterns, we go a step forward and assume that semantic structures describing concept knowledge can be mapped on sets of lexico-syntactic patterns.

5 Conclusions

In this paper we presented a novel method for unsupervised knowledge extraction for taxonomies of concepts using Wikipedia as information source. Departing from previous methods for knowledge acquisition we seek to extract semantic structures from wikipedia descriptions of similar concepts. These structures are formalized as surface patterns linking the title concepts with their properties. Future work includes:

1. usage of more formalized taxonomies.
2. the extension of the set of taxonomies to include abstract concepts like **cognition**.
3. a better evaluation framework for the results.

Acknowledgments

The authors would like to thank to Verginica Barbu Mititelu and Gianluca Lebani for support in the data collection and data rating. We also want to thank three anonymous reviewers for suggestions and criticism.

References

[1] M. A. Hearst. Automated discovery of wordnet relations. In C. Fellbaum, editor, *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[2] A. Herbelot and A. Copestake. Acquiring ontological relationships from wikipedia using rmrs. In *ISWC Workshop On Web Content*, 2006.

[3] D. N. Milne, I. H. Witten, and D. M.Nichols. A knowledge-based search engine powered by wikipedia. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 445–454, New York, NY, USA, 2007. ACM.

[4] M.Pasca and B. Durme. Weakly-supervised acquisition of open-domain classes and class knowledge from web documents and query logs. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2008.

[5] S. Pradhan, E. Loper, D. Dligach, and M. Palmer. Semeval-2007 task-17: English lexical sample, srl and all words. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 87–92, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[6] M. Ruiz-Casado, E. Alfonseca, and P. Castells. Automatising the learning of lexical patterns: An application to the enrichment of wordnet by extracting semantic relationships from wikipedia. *Data and Knowledge Engineering*, 61(3):484–499, 2007. Advances on Natural Language Processing - NLDB 05.

[7] S. Siegel and N. J. Castellan. *Nonparametric statistics for the Behavioral Sciences*. McGraw-Hill, 2nd edition, 1988.

[8] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: a core of semantic knowledge. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 697–706, New York, NY, USA, 2007. ACM.

[9] G. Wang, Y. Yu, and H. Zhu. Positive-only relation extraction from wikipedia text. In *Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference, ISWC/ASWC'07*, 2007.