Large-Scale Taxonomy Mapping for Restructuring and Integrating Wikipedia

Simone Paolo Ponzetto
Seminar für Computerlinguistik
University of Heidelberg
ponzetto@cl.uni-heidelberg.de

Roberto Navigli
Dipartimento di Informatica
Università di Roma "La Sapienza"
navigli@di.uniroma1.it

Abstract

We present a knowledge-rich methodology for disambiguating Wikipedia categories with WordNet synsets and using this semantic information to restructure a taxonomy automatically generated from the Wikipedia system of categories. We evaluate against a manual gold standard and show that both category disambiguation and taxonomy restructuring perform with high accuracy. Besides, we assess these methods on automatically generated datasets and show that we are able to effectively enrich WordNet with a large number of instances from Wikipedia. Our approach produces an integrated resource, thus bringing together the fine-grained classification of instances in Wikipedia and a well-structured top-level taxonomy from WordNet.

1 Introduction

The need of structured knowledge for intelligent systems is a *leitmotiv* of Artificial Intelligence (AI) – starting from [McCarthy, 1959] till current echoes in [Schubert, 2006]. Previous efforts aiming at maximizing the quality of knowledge repositories have concentrated on collecting this knowledge manually: the WordNet project [Fellbaum, 1998] for instance provides a semantic lexicon for English and has become *de facto* the most widely used knowledge resource in Natural Language Processing (NLP). However, while providing a comprehensive repository of word senses, WordNet contains very little domain-oriented knowledge and is populated with only a few thousand instances, i.e. named entities.

To overcome the limitations of manually assembled knowledge repositories, research efforts in AI and NLP have been devoted to automatically harvest that knowledge [Buitelaar *et al.*, 2005]. In particular, the last years have seen a growing interest for the automatic acquisition of machine readable knowledge from semi-structured knowledge repositories such as Wikipedia [Suchanek *et al.*, 2007; Nastase and Strube, 2008; Wu and Weld, 2008, inter alia]. Nonetheless, questions remain whether these automatically-induced knowledge resources achieve the same quality of manually engineered ones, such as WordNet or Cyc [Lenat and Guha, 1990].

The most notable strength of Wikipedia, i.e. its very large coverage, lies not only in its large number of encyclopedic entries, but also in the domain orientation of its categorization network, i.e. very specific categories such as MEDICINAL PLANTS or WOODLAND SALAMANDERS[1]. However, such categorization system is merely a thematically organized thesaurus. Although methods have been developed to induce taxonomies from it [Ponzetto and Strube, 2007, WikiTaxonomy henceforth], these cope badly with very general concepts. This is because the upper regions of the Wikipedia categorization are almost exclusively thematic, and no subsumption relation can be found while remaining inside the category network. For instance, COUNTRIES is categorized under (*isa*) PLACES, which in turn is categorized under GEOGRAPHY and NATURE, thus having no suitable parent dominating it with a subsumption relation. This is reflected in the 3,487 roots included in WikiTaxonomy: the resource is a sparse set of taxonomic islands in need to be linked to more general concepts for it to resemble a sane taxonomy. In addition, being automatically generated, manual inspection of that resource reveals several errors, e.g. FRUITS *isa* PLANTS, which can be automatically corrected by enforcing taxonomic constraints from a reference ontology, i.e. given a taxonomy mapping, one could recover from errors by aligning the automatically generated taxonomy to a manual one.

We tackle these issues by proposing a two-phase methodology. The method starts with WikiTaxonomy, although in principle any taxonomy can be input. In a first step, the taxonomy is automatically mapped to WordNet. This mapping can be cast as a Word Sense Disambiguation (WSD) problem [Navigli, 2009]: given a Wikipedia category (e.g. PLANTS), the objective is to find the WordNet synset that best captures the meaning of the category label (e.g. plant^2_n)[2]. The optimal mapping is found based on a knowledge-rich method which maximizes the structural overlap between the source and target knowledge resources. As a result, the Wikipedia taxonomy is automatically 'ontologized'. Secondly, the mapping outcome of the first phase is used to restructure the Wikipedia taxonomy itself. Restructuring operations are applied to those Wikipedia categories which convey the highest degree of inconsistency with respect to the corresponding part

[1] We use Sans Serif for words, CAPITALS for Wikipedia pages and SMALL CAPS for Wikipedia categories.

[2] We denote with w^i_p the i-th sense of a word w with part of speech p. We use word senses to unambiguously denote the corresponding synsets (e.g. plant^2_n for { plant, flora, plant life }).

of the WordNet subsumption hierarchy. This ensures that the structure of the Wikipedia taxonomy better complies with a reference manual resource. In fact, category disambiguation and taxonomy restructuring synergetically profit from each other: disambiguated categories allow it to enforce taxonomic constraints and a restructured taxonomy in turn provides a better context for category disambiguation.

Our approach to taxonomy mapping and restructuring provides three contributions: first, it represents a sound and effective methodology for enhancing the quality of an automatically extracted Wikipedia taxonomy; second, as an additional outcome, we are able to populate a reference taxonomy such as WordNet with a large amount of instances from Wikipedia; finally, by linking WikiTaxonomy to WordNet we create a new subsumption hierarchy which includes in its lowest regions the fine-grained classification from Wikipedia, and in its upper regions the better structured content from WordNet. This allows to connect the taxonomic islands found in WikiTaxonomy via WordNet, since the higher regions of the merged resource are provided by the latter.

2 Methodology

Our methodology takes as input a Wikipedia taxonomy (Section 2.1). First, it associates a synset with each Wikipedia category in the taxonomy (Section 2.2). Next, it restructures the taxonomy in order to increase its alignment with the WordNet subsumption hierarchy (Section 2.3).

2.1 Preliminaries

We take as input WikiTaxonomy[3]. We can view the taxonomy as a forest $\mathcal{F}$ of category trees. As an example, in Figure 1 we show an excerpt of the category tree rooted at PLANTS. Each vertex in the tree represents a Wikipedia category. The label of this category is often a complex phrase, e.g. JAZZ HARMONICA PLAYERS BY NATIONALITY. In order to produce a mapping to WordNet, we need to find the lexical items $heads(c)$ best matching each category label c, e.g. JAZZ HARMONICA PLAYERS can be mapped to any WordNet sense of player. Terms in WordNet (we use version 3.0) are first searched for a full match with the category label, e.g. plant for PLANTS. If no full match is found, we fall back to the *head* of the category. First, the lexical heads of a category label are found using a state-of-the-art parser [Klein and Manning, 2003]. Then, we take as head of a category the minimal NP projection of its lexical head, e.g. public transport for PUBLIC TRANSPORT IN GERMANY. Such NP is found in the parse tree by taking the head terminal and percolating up the tree until the first NP node is found. If no such minimal NP can be found in WordNet, we take the lexical head itself, e.g. plants for EDIBLE PLANTS. In case of coordinations we collect both lexical heads, e.g. building and structure for BUILDINGS AND STRUCTURES IN GERMANY.

2.2 Category Disambiguation

For each category tree $T \in \mathcal{F}$ and for each category $c \in T$, we first produce a mapping from c to the most appropriate synset $\mu_T(c)$ – e.g. we want to associate plant^2_n (the botany sense) with category PLANTS from Figure 1. Category disambiguation is performed in two steps:

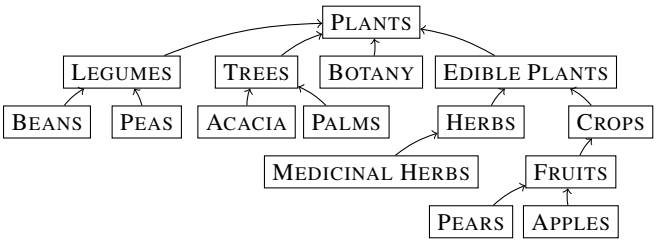

Figure 1: An excerpt of the Wikipedia category tree rooted at PLANTS.

1. **WordNet graph construction.** We start with an empty graph $G = (V, E)$. For each category $c \in T$, and for each head $h \in heads(c)$, the set of synsets containing h is added to V. For instance, given the category BEANS we add to V the synsets which contain the four WordNet senses of bean (namely, 'edible seed', 'similar-to-bean seed', 'plant', and 'human head'). Next, for each vertex $v_0 \in V$ we set $v = v_0$ and we climb up the WordNet *isa* hierarchy until either we reach its root or we encounter a vertex $v' \in V$ (e.g. legume^1_n is a parent of bean^3_n). In the latter case, if $(v, v') \notin E$ we add it to E (e.g. we add (bean^3_n, legume^1_n) to E) and set its weight $w(v, v')$ to 0. Finally, for each category $c' \in T$ whose head occurs in the synset v' (in our example, LEGUMES), the edge weight $w(v, v')$ is increased as follows:

$$w(v, v') = w(v, v') + \frac{1}{2^{d_{WN}(v_0, v')-1} \cdot 2^{d_{Wiki}(c_0, c')-1}}$$

where $d_{WN}(v_0, v')$ is the number of subsumption edges between v_0 and v' in WordNet and $d_{Wiki}(c_0, c')$ is the number of edges between c_0 (the category corresponding to v_0) and c' in the category tree T (set to the depth D of our tree if c' is not an ancestor of c_0). The procedure is repeated iteratively by setting $v = v'$, until the root of the WordNet hierarchy is reached. In our example, we have that d_{WN}(bean^3_n, legume^1_n) = 1 and d_{Wiki}(BEANS, LEGUMES) = 1, thus the weight of the corresponding edge is set to $1/(2^{1-1} \cdot 2^{1-1}) = 1$. Analogously, we update the weights on the path legume$^1_n \rightarrow \cdots \rightarrow$ plant^2_n. We note that the contribution added to $w(v, v')$ exponentially decreases with the distance between v_0 and v' and between c_0 and c'. At the end of this step, we obtain a graph G including all possible sense interpretations of all categories in our tree T. In Figure 2 we show an excerpt of the WordNet graph associated with the category tree of Figure 1.

2. **Disambiguation.** As a second step, we use the resulting WordNet graph to identify the most relevant synset for each Wikipedia category $c \in T$. First, the set of edges E is sorted in decreasing order according to the edges' weight. For each edge (v, v') in such ordered set, if the corresponding category c (c') has not been assigned a synset before, we set $\mu_T(c) = v$ ($\mu_T(c') = v'$). For instance, in E we have (tree^1_n, plant^2_n) with weight 0.37 and lower weights

[3] www.eml-research.de/nlp/download/wikitaxonomy.php

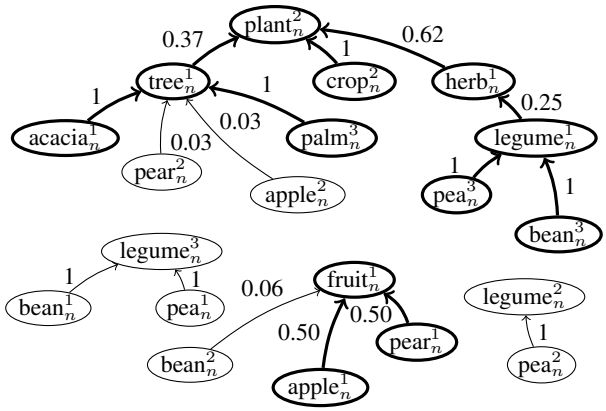

Figure 2: An excerpt of the WordNet graph associated with the category tree rooted at PLANTS. Thick lines correspond to highest-ranking edges and their incident vertices selected as sense interpretations for the corresponding categories. Singleton vertices are not shown.

for edges involving other senses of tree and plant, thus we assign μ_T(TREES) = tree^1_n and μ_T(PLANTS) = plant^2_n. We repeat this step until we assign a synset to each category in our tree or the entire list of edges is examined. In case of edges with the same maximum weight which connect different senses for the same category c, we assign it the synset v which maximizes the size of the connected component of G it belongs to. In Figure 2 thick lines highlight high-ranking edges and the incident synsets chosen as sense interpretations of the respective categories.

2.3 Taxonomy Restructuring

The second phase of our methodology performs a restructuring of our category forest $\mathcal{F}$ which aims at increasing its degree of alignment to the reference taxonomy (i.e. WordNet). For each tree $T \in \mathcal{F}$ we perform the following three steps:

1. **Edge penalty weighting.** For each edge $e \in T$, we set the initial penalty of e to 0, i.e. $p(e) = 0$. Next, for each vertex $c_0 \in T$, we analyze the path $c_0 \rightarrow c_1 \rightarrow \cdots \rightarrow c_n$ leading from c_0 to the root c_n of T. For each edge (c_i, c_{i+1}) along the path, if there is no subsumption path in WordNet between the senses assigned to c_0 and c_{i+1} (i.e. we cannot infer that $\mu_T(c_0)$ *isa* $\mu_T(c_{i+1})$), we set:

$$p(c_i, c_{i+1}) = p(c_i, c_{i+1}) + \frac{1}{2^{d_{Wiki}(c_0, c_{i+1})-1}}$$

i.e. we increase the penalty of edge (c_i, c_{i+1}) by an inverse exponential function of the distance between c_0 and c_{i+1} in the category tree T. For instance, consider c_0 = FRUITS. Its path to the root is: FRUITS $\rightarrow$ CROPS $\rightarrow$ EDIBLE PLANTS $\rightarrow$ PLANTS. However, there is no *isa* relation in WordNet between fruit^1_n and crop^2_n (the senses assigned to the corresponding categories), thus the penalty p(FRUITS, CROPS) is increased by $1/2^0 = 1$. Similarly, no *isa* relation holds in WordNet between fruit^1_n and plant^2_n, so the penalty p(CROPS, EDIBLE PLANTS) is increased by $1/2^1 = 0.5$ and p(EDIBLE PLANTS, PLANTS) is increased by $1/2^2 = 0.25$.

2. **Identification of maximum penalty cuts.** To identify those edges in T which maximize the penalty value (and thus the degree of inconsistency compared to other vertices in T), we sort the set of edges by penalty and select the subset P_α which includes the top α percentage of them. In our example, our set P_α includes the following edges { (BOTANY, PLANTS), (FRUITS, CROPS), (LEGUMES, PLANTS) } for $\alpha = 0.3$[4].

3. **Tree restructuring.** Finally, for each high-penalty edge $(c, c') \in P_\alpha$, we determine whether we can find a better attachment for c within the entire forest $\mathcal{F}$. If exists a category $c'' \in T'$ ($T' \in \mathcal{F}$) for which a direct subsumption relation can be identified in WordNet between the senses assigned to c and c'', i.e. $\mu_T(c)$ *isa* $\mu_{T'}(c'')$, then we move c under c'' (that is we remove (c, c') from T and add (c, c'') to T'). Note that it is not necessarily true that $T' \neq T$. For example, given the edge (LEGUMES, PLANTS) we find that legume^1_n *isa* herb^1_n in WordNet, thus we can move the subtree rooted at LEGUMES under vertex HERBS. Similarly BOTANY is moved under BIOLOGY (a vertex whose ancestor is SCIENCE). However, FRUITS cannot be moved, as the direct hypernym of fruit^1_n (the WordNet sense assigned to FRUITS) is reproductive_structure^1_n, which does not have a Wikipedia counterpart.

3 Evaluation

In order to evaluate our methodology, we quantify its performance with respect to its two phases:

1. **Category disambiguation.** How good is the system at selecting the correct WordNet senses for the Wikipedia category labels?
2. **Taxonomy restructuring.** How good is the restructuring of the taxonomy based on the disambiguated categories?

To do so, we perform a manual evaluation against manually-tagged datasets for the two tasks (Sections 3.1 and 3.2). In addition, we devise an automated assessment against datasets constructed with the aid of Wikipedia instances mapped to monosemous WordNet senses (Section 3.3).

3.1 Category disambiguation: manual evaluation

In order to create a gold standard for category disambiguation, we randomly sampled 2,000 categories from Wikipedia. These were manually annotated with WordNet synsets by one annotator with previous experience in lexicographic annotation. Each category was paired with the appropriate WordNet synset for its lexical head (e.g. theatre^1_n for THEATRES IN AUSTRIA and theatre^2_n for THEATRE IN SCOTLAND). In order to quantify the quality of the annotations and the difficulty of the task, a second annotator was asked to sense tag the 310 categories with the five most frequent lexical heads

[4]The optimal value for α was found based on a random sample of trees containing 10% of the categories in the taxonomy.

	tree size			
	2-9	10-100	>100	overall
category disambiguation	62.1	77.7	81.5	80.8
random baseline	36.3	44.2	46.6	46.3
most frequent sense	60.4	69.0	75.2	74.5
# trees	9	65	133	207

Table 1: Category disambiguation: manual evaluation

from the dataset and the inter-annotator agreement using the kappa coefficient [Carletta, 1996] was computed. Our annotators achieved an agreement coefficient κ of 0.92 indicating almost perfect agreement.

Table 1 shows the accuracy of our disambiguation algorithm against the manually annotated categories. For each Wikipedia category we evaluate whether it has been mapped to the correct WordNet sense. For those categories which cannot be disambiguated we select the most frequent sense (i.e. the first sense listed in the WordNet sense inventory). As a baseline we take the most frequent sense for all categories and the random baseline as a lower bound. Since our method pivotally relies on the size of the local category tree to be disambiguated, i.e. since it takes the ancestors and descendants of a category into account as its context for disambiguation, we additionally evaluate for trees of different sizes and report the average accuracy of their disambiguation.

3.2 Taxonomy restructuring: manual evaluation

To assess the taxonomy restructuring procedure, we manually evaluated the accuracy of the single restructuring steps, e.g. if the link between LEGUMES and PLANTS is removed and LEGUMES is moved under HERBS, we check the correctness of the operation. We selected a random sample of 200 moves, i.e. detachment-attachment pairs for evaluation. Each detachment-attachment edge pair (d, a) was evaluated by two annotators as follows: correct, if either i) the original edge d was incorrect and the newly added link a was correct, ii) d was correct and a specializes d; incorrect, otherwise. Again, we computed the inter-annotator agreement for the sample ($\kappa = 0.75$) and removed those pairs for which there was no agreement (12 in total). On this manually annotated dataset we achieve an accuracy of 88.8% for the task of taxonomy restructuring. Examples of good restructuring operations are SUPERMARKET TABLOIDS moved from NATIONAL NEWSPAPERS PUBLISHED IN THE UNITED STATES to NEWSPAPERS, or ARISTOTLE moved from CLASSICAL GREEK PHILOSOPHY to PHILOSOPHERS. Incorrect restructuring operations are determined by previous errors in the disambiguation step, e.g. MANHATTAN moved from NEW YORK COUNTIES to COCKTAILS, or HAMILTON, ONTARIO moved from CITIES IN ONTARIO to MATHEMATICIANS.

3.3 Instance-based automated evaluation

The evaluations in Sections 3.1 and 3.2 were performed against manual gold standards. To perform a large-scale assessment, we devised a method to automatically produce two evaluation datasets D and D' for taxonomy disambiguation and taxonomy consistency, respectively. The datasets are obtained as a result of the following two steps.

Instance collection. Given a category tree T, for each category $c \in T$ in Wikipedia we first collect its instances. The instances of a category are automatically found by using the heuristics from YAGO [Suchanek *et al.*, 2007]: we collect all pages for each category whose label contains a plural lexical head, e.g. AMPHIUMA *instance-of* SALAMANDERS. In addition, in order to filter incorrect instance assignments, e.g. XYLOTHEQUE *instance-of* BOTANICAL GARDENS, we check whether the page title occurs in HeiNER, a gazetteer automatically induced from Wikipedia [Wentland *et al.*, 2008] based on the heuristics developed in [Bunescu and Paşca, 2006]. Finally, we filter out instances that do not occur or are not monosemous in WordNet.

Dataset construction. Given a Wikipedia instance i of a category c (e.g. AMPHIUMA is an instance of SALAMANDERS), and given its corresponding WordNet synset $S_{c,i}$ (e.g. amphiuma^1_n corresponds to AMPHIUMA), we identify those WordNet ancestors $S_{c',i}$ of $S_{c,i}$ such that some Wikipedia category c' maps to it (e.g. amphibian^3_n is an ancestor of amphiuma^1_n, corresponding to category AMPHIBIANS). For each such ancestor $S_{c',i}$, we populate two datasets as follows:

- we add to our category disambiguation dataset D the pair $(c', S_{c',i})$, thus implying that the correct sense for c' is $S_{c',i}$ (e.g. we add (AMPHIBIANS, amphibian^3_n));
- we add to our taxonomy consistency dataset D' the *isa* pair (c, c') (e.g. (SALAMANDERS, AMPHIBIANS)).

In other words, for category disambiguation we exploit instances to identify synsets $S_{c',i}$ whose lexical items correspond to categories c' in T, whereas for taxonomy consistency we identify pairs of categories (c, c') in *isa* relation based on a corresponding WordNet path connecting synset $S_{c,i}$ to $S_{c',i}$. Dataset construction is based on instances since these are collected from Wikipedia pages which are unseen to the disambiguation and restructuring phases of our method. Furthermore, we consider only monosemous instances, since these have a univocal WordNet mapping. This allows us to use the WordNet subsumption hierarchy as an oracle for sense tagging categories corresponding to the instances' ancestors, and for identifying subsumption relationships between categories. This evaluation methodology automatically generates gold-standard mappings between Wikipedia categories and WordNet synsets, that is, it provides a method for automatically disambiguating Wikipedia categories. However, it achieves very low coverage – i.e. it disambiguates only 17.3% of the categories in WikiTaxonomy – and it is used only for evaluation purposes.

We determined the accuracy of our category disambiguation phase against our dataset D before and after the restructuring phase. Similarly, we calculated the performance of taxonomy restructuring by checking the taxonomy consistency against dataset D' before and after restructuring. As in the case of the manual evaluation, we take the most frequent sense as a baseline for category disambiguation together with the random baseline. The results are reported in Table 2.

	before restructuring	after restructuring
category disambiguation	95.3	95.7
random baseline	63.1	63.1
most frequent sense	79.1	78.5
taxonomy consistency	38.4	44.3
# test instances	70,841	73,490

Table 2: Results on instance-based evaluation

3.4 Discussion

Our methodology achieves good results in the manual and automatic evaluation for both disambiguation and restructuring tasks. On the task of category disambiguation we achieve an improvement of $+6.3\%$ and $+16.2\%$ with respect to the most frequent sense baseline in the manual and automatic evaluation respectively. Taxonomy restructuring improves the performance on the instance-based evaluation of the category disambiguation ($+0.4\%$), and achieves $+5.9\%$ on the taxonomy consistency evaluation, i.e. quantifying the degree of alignment of the Wikipedia taxonomy to WordNet.

In general, the task of category disambiguation seems beneficial for a number of other steps. First, it allows to link instances from Wikipedia pages to WordNet synsets. Using the WordNet subsumption hierarchy as gold standard allows in turn to automatically generate a large dataset of semantically classified named entities at practically no cost[5]. In this respect, our results suggest a methodology for a more accurate population of WordNet with instances than YAGO, which only relies on the most frequent sense heuristic. Merely relying on the most frequent WordNet sense for mapping a Wikipedia category does not in fact seem to suffice, given that the average polysemy of Wikipedia categories is 3.2 senses per category. For instance, in our initial example of Figure 1, mapping PLANTS to plant_n^1 would imply to take all instances under this category in Wikipedia as instances of industrial plants. As a result, one reduces one of the most appealing strengths of Wikipedia, i.e. its large-coverage at the instance level. In this respect, our work is similar in spirit to that of [Snow *et al.*, 2006]: extending WordNet with new information, i.e. named entities, while enforcing constraints which model lexical ambiguity.

In addition, the disambiguation allows to restructure WikiTaxonomy, thus recovering from errors induced during its automatic acquisition. By moving categories to more suitable parents, as highlighted by the high accuracy obtained in Section 3.2, not only we generate an improved taxonomy with more correct *isa* relations, but also a better context for disambiguation, as shown in the overall improvement in Table 2. Quantitative analysis of the restructured taxonomy reveals trees of an average smaller size (95.8 versus 75.6) and depth (2.9 versus 2.8), thus suggesting that rather than relying only on the tree size, the category disambiguation pivotally relies on its quality, since the trees themselves provide the context for disambiguation.

[5]The mappings from Wikipedia categories to WordNet synsets can be freely downloaded at `www.cl.uni-heidelberg.de/~ponzetto/wikitax2wn`.

Finally, linking Wikipedia categories to WordNet synsets allows it to overcome the sparseness of WikiTaxonomy. For instance, given our mapping for PLANTS, one can use the WordNet hierarchy to leverage the additional information that plant_n^2 *isa* organism_n^1, e.g. in order to compute the semantic similarity between ANIMALS and PLANTS even if these two categories belong to different trees in WikiTaxonomy.

4 Related work

An approach to map Wikipedia to WordNet is first presented in [Ruiz-Casado *et al.*, 2005], where each Wikipedia page is assigned to its most similar WordNet synset, based on the similarity between the page's text and synsets' glosses in a Vector Space Model. This method relies only on text overlap techniques and does not take advantage of the input from Wikipedia being semi-structured.

Recent years have witnessed a considerable amount of work in information extraction to generate structured semantic content from Wikipedia. [Ponzetto and Strube, 2007] present a set of lightweight heuristics to generate a large-scale taxonomy from the system of Wikipedia categories. Their methods automatically assign *isa* and *not-isa* labels to the relations between categories: extensions have been later proposed to refine the *isa* relation to classify categories as *instances* or *classes* [Zirn *et al.*, 2008], as well as to generate more specific relations from the *not-isa* relations, e.g. *part-of*, *located-in*, etc. [Nastase and Strube, 2008].

The problem of generating an ontology from Wikipedia and mapping it to WordNet is taken on by the YAGO [Suchanek *et al.*, 2007] and Kylin Ontology Generator [Wu and Weld, 2008, KOG] systems. YAGO merges WordNet and Wikipedia by populating the subsumption hierarchy of WordNet with instances taken from Wikipedia pages. Instances (described by Wikipedia pages) are first assigned an *instance-of* relation to categories if the head of the page's category label is plural, e.g. ANGELICA SINENSIS is an instance of the concept denoted by MEDICINAL PLANTS. Wikipedia categories are then mapped to WordNet by selecting the synset which contains the most frequent sense of their label (or lexical head) – e.g. PLANTS and MEDICINAL PLANTS both get mapped to plant_n^1 as in 'industrial plant'. The heuristic for instance-class assignment is as simple as high performing, and we use it to collect the instances for our instance-based evaluation (Section 3.3). However, we also move away from the most frequent sense heuristic and merely use it as a baseline. KOG builds a subsumption hierarchy of classes by combining Wikipedia infoboxes with WordNet using statistical-relational learning. Each infobox template, e.g. `Infobox Plants` for plants, represents a class and the slots of the template are considered as the attributes of the class. While KOG represents in many ways the theoretically soundest methodology to induce a taxonomy from the structure of Wikipedia, i.e. by jointly predicting both the subsumption relation between classes and their mapping to WordNet, it still uses the most frequent sense heuristic in case multiple senses are available for the same infobox class.

Approaches that go beyond the most frequent sense heuristic (including ours) view the mapping of Wikipedia to Word-

Net as a sense disambiguation problem. These include methods based on classical WSD algorithms [Mihalcea, 2007] as well as cross-referencing documents with encyclopedic repositories [Milne and Witten, 2008]. In this light, our goal is slightly different, in that we do not sense-tag words occurring within a Wikipedia page, but categories structured in a subsumption hierarchy, similar to what [Navigli and Velardi, 2004] do with domain terminology. However, the latter approach is local, in that it is restricted to term trees organized by string inclusion, whereas in our approach we disambiguate and restructure a full-fledged conceptual taxonomy.

5 Conclusions

In this paper we proposed a novel method for restructuring and integrating a Wikipedia taxonomy. Key to our approach is the use of a reference manual resource, namely the WordNet subsumption hierarchy, to perform category disambiguation and taxonomy restructuring. Both phases are performed with high accuracy as experimentally determined through manual and automatic evaluations. Our method leads to the integration of Wikipedia and WordNet, in that Wikipedia categories are enriched with accurate sense information from WordNet and WordNet synsets are effectively populated with instances from Wikipedia.

Our approach is resource-independent and can be applied to improve any automatically-acquired taxonomy with the aid of a manually-constructed one. For instance, the approach can be easily adapted to integrate Wikipedia with other reference resources such as Cyc, as long as these provide a subsumption hierarchy (in contrast, e.g., with [Medelyan and Legg, 2008], whose method is tailored on Cyc features).

While our methodology is automatic, it can be used in a semi-automatic manner so as to allow human users to perform computer-assisted taxonomy validation. In fact, our taxonomy restructuring operations can be presented to the user sorted by penalty score, thus minimizing the human effort by presenting the more reliable moves first.

As future work, we plan to employ the integrated knowledge base obtained from our method to perform open-text Word Sense Disambiguation as well as knowledge-lean Question Answering.

References

[Buitelaar *et al.*, 2005] Paul Buitelaar, Philipp Cimiano, and Bernardo Magnini, editors. *Ontology Learning from Text: Methods, Evaluation and Applications*. Amsterdam, The Netherlands: IOS Press, 2005.

[Bunescu and Paşca, 2006] Razvan Bunescu and Marius Paşca. Using encyclopedic knowledge for named entity disambiguation. In *Proc. of EACL-06*, pages 9–16, 2006.

[Carletta, 1996] Jean Carletta. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254, 1996.

[Fellbaum, 1998] Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, Mass., 1998.

[Klein and Manning, 2003] Dan Klein and Christopher D. Manning. Accurate unlexicalized parsing. In *Proc. of ACL-03*, pages 423–430, 2003.

[Lenat and Guha, 1990] Douglas B. Lenat and R. V. Guha. *Building Large Knowledge-Based Systems: Representation and Inference in the CYC Project*. Addison-Wesley, Reading, Mass., 1990.

[McCarthy, 1959] John McCarthy. Programs with common sense. In *Proc. of the Teddington Conference on the Mechanization of Thought Processes*, pages 75–91. London, U.K., 1959.

[Medelyan and Legg, 2008] Olena Medelyan and Catherine Legg. Integrating cyc and wikipedia: Folksonomy meets rigorously defined common-sense. In *Proc. of the AAAI Workshop on Wikipedia and Artificial Intelligence*, pages 13–18, Chicago, Ill., 2008.

[Mihalcea, 2007] Rada Mihalcea. Using Wikipedia for automatic word sense disambiguation. In *Proc. of NAACL-HLT-07*, pages 196–203, 2007.

[Milne and Witten, 2008] David Milne and Ian H. Witten. Learning to link with Wikipedia. In *Proc. of CIKM-08*, pages 509–518, 2008.

[Nastase and Strube, 2008] Vivi Nastase and Michael Strube. Decoding Wikipedia category names for knowledge acquisition. In *Proc. of AAAI-08*, pages 1219–1224, 2008.

[Navigli and Velardi, 2004] Roberto Navigli and Paola Velardi. Learning domain ontologies from document warehouses and dedicated websites. *Computational Linguistics*, 30(2):151–179, 2004.

[Navigli, 2009] Roberto Navigli. Word Sense Disambiguation: a survey. *ACM Computing Surveys*, 41(2):1–69, 2009.

[Ponzetto and Strube, 2007] Simone Paolo Ponzetto and Michael Strube. Deriving a large scale taxonomy from Wikipedia. In *Proc. of AAAI-07*, pages 1440–1445, 2007.

[Ruiz-Casado *et al.*, 2005] Maria Ruiz-Casado, Enrique Alfonseca, and Pablo Castells. Automatic assignment of Wikipedia encyclopedic entries to WordNet synsets. In *Advances in Web Intelligence*, volume 3528 of *Lecture Notes in Computer Science*. Springer Verlag, 2005.

[Schubert, 2006] Lenhart K. Schubert. Turing's dream and the knowledge challenge. In *Proc. of AAAI-06*, pages 1534–1538, 2006.

[Snow *et al.*, 2006] Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. Semantic taxonomy induction from heterogeneous evidence. In *Proc. of COLING-ACL-06*, pages 801–808, 2006.

[Suchanek *et al.*, 2007] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. YAGO: A core of semantic knowledge. In *Proc. of WWW-07*, pages 697–706, 2007.

[Wentland *et al.*, 2008] Wolodja Wentland, Johannes Knopp, Carina Silberer, and Matthias Hartung. Building a multilingual lexical resource for named entity disambiguation, translation and transliteration. In *Proc. of LREC '08*, 2008.

[Wu and Weld, 2008] Fei Wu and Daniel Weld. Automatically refining the Wikipedia infobox ontology. In *Proc. of WWW-08*, 2008.

[Zirn *et al.*, 2008] Cäcilia Zirn, Vivi Nastase, and Michael Strube. Distinguishing between instances and classes in the Wikipedia taxonomy. In *Proc. of ESWC-08*, pages 376–387, 2008.