Revisiting Taxonomy Induction over Wikipedia

Amit Gupta
EPFL
Lausanne, Switzerland
amit.gupta@epfl.ch

Francesco Piccinno
University of Pisa
Pisa, Italy
piccinno@di.unipi.it

Mikhail Kozhevnikov
Google Inc.
Zurich, Switzerland
qnan@google.com

Marius Paşca
Google Inc.
Mountain View, California
mars@google.com

Daniele Pighin
Google Inc.
Zurich, Switzerland
biondo@google.com

Abstract

Guided by multiple heuristics, a unified taxonomy of entities and categories is distilled from the Wikipedia category network. A comprehensive evaluation, based on the analysis of upward generalization paths, demonstrates that the taxonomy supports generalizations which are more than twice as accurate as the state of the art. The taxonomy is available at http://headstaxonomy.com.

1 Introduction

Motivation. As possibly the largest resource of publicly available, semi-structured knowledge, Wikipedia (Remy, 2002) serves as a stepping stone towards the construction of collections of structured data (Remy, 2002; Hoffart et al., 2013; Vrandečić and Krötzsch, 2014). Data within Wikipedia benefits from new additions and distributed curation by human editors, and has proven beneficial in text analysis tasks ranging from co-reference resolution (Ratinov and Roth, 2012), word sense (Mihalcea, 2007) and entity disambiguation (Ratinov et al., 2011), to information retrieval (Hu et al., 2009) and information extraction (Wu and Weld, 2010; Nastase and Strube, 2013; Hoffart et al., 2013; Dong et al., 2014).

Wikipedia links millions of entities (e.g. *Barack Obama*) to thousands of inter-connected categories of different granularity (e.g. *Presidents of the United States*, *Political office-holders, Politicians*) to form what is often referred to as the Wikipedia category network (WCN). However, obtaining a taxonomy of increasingly general categories from WCN is by no means trivial because upward edges in WCN, from entities to categories and also from child to parent categories, are not confined to *is-a* relations (Ponzetto and Strube, 2007). In fact, consistently discarding *not-is-a* edges such as *Japan*$\rightsquigarrow$*660 BC* or *Award winners*$\rightsquigarrow$*Awards*, while retaining as many true *is-a* edges as possible, has been the object of a steady body of research (Ponzetto and Strube, 2007; Hovy et al., 2013; Flati et al., 2014). Recent methods still produce taxonomies with glaring gaps in precision and coverage. More importantly, even if the methods correctly identify individual *is-a* edges with an accuracy as high as 85% (Flati et al., 2014), it is not uncommon for upward paths to traverse at least some incorrect edges. The resulting taxonomies transitively connect entities (e.g., *Natural language processing*) to many ancestor categories (e.g., *Physical body*, *Mass*)[1] that are not true generalizations, thus limiting their utility in practice.

Contributions. This paper proposes a novel method for taxonomy induction from WCN. As described in Section 3, the method exploits syntactic evidence in category titles to connect entities (i.e., pages) with increasingly more general categories. A novel, comprehensive framework for taxonomy evaluation is proposed, focusing on the accuracy and granularity of longer generalization paths, as opposed to individual edges. Section 4 describes the evaluation framework and carries out an in-depth comparison of the proposed taxonomy against the state of the art. It shows significant gains in accuracy relative to current state of the art, while maintaining similar coverage.

[1] Examples taken from http://wibitaxonomy.org .

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 2300–2309, Osaka, Japan, December 11-17 2016.

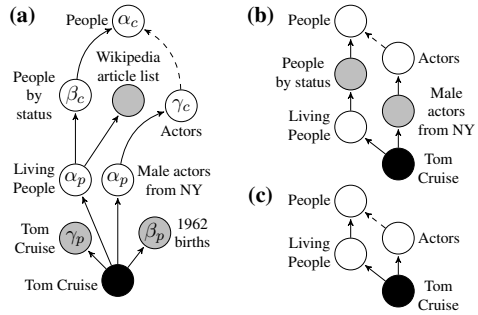

Figure 1: Taxonomy induction phases. Black circles denote entities. White circles denote categories. Dashed lines denote paths including possibly multiple edges. **(a)** Step 1: Page heuristics (α_p, β_p and γ_p) and category heuristics (α_c, β_c and γ_c) are applied sequentially to select candidate generalizations for each node (page or category), until one produces at least one candidate (white circles). Gray nodes show candidates that would have been produced by remaining heuristics. **(b)** Step 2: Nodes that encode redundant information are removed (grey). **(c)** Resulting taxonomy.

2 Related work

Thanks to continuous contributions and curation by many human editors, Wikipedia (Remy, 2002) recommends itself as a high quality resource of semi-structured knowledge. It enables multiple approaches to large scale knowledge acquisition and taxonomy induction (Hovy et al., 2013). One of the earliest attempts towards the latter is WikiTaxonomy (Ponzetto and Strube, 2007; Ponzetto and Strube, 2011). In WikiTaxonomy, relations are labeled as either *is-a* or *not-is-a*, using a cascade of heuristics based on the syntactic structure of category labels, the topology of the network and lexico-syntactic patterns for detecting subsumption and meronymy, similar to Hearst patterns (Hearst, 1992). WikiNet (Nastase et al., 2010) extends WikiTaxonomy by expanding *not-is-a* relations into fine-grained relations such as *part-of, located-in, etc*. YAGO, induces a taxonomy by employing heuristics linking Wikipedia categories to corresponding synsets in WordNet (Hoffart et al., 2013). YAGO's taxonomy forms the backbone of a variety of intelligent applications, including Watson (Ferrucci et al., 2010). DBPedia (Lehmann et al., 2015) aims to provide a fully-structured representation of semi-structured content of Wikipedia. It focuses on linking the extracted knowledge with existing resources such as YAGO, OpenCyc etc.

The Wikipedia Bitaxonomy project, or WIBI (Flati et al., 2014), the most recent effort towards large-scale taxonomy induction from Wikipedia, simultaneously induces a taxonomy for pages and a separate taxonomy for categories from WCN using the idea that information contained in pages can be useful in constructing a taxonomy of categories and vice-versa. First, an initial taxonomy over pages is constructed by extracting lemmas from their first sentences and resolving them to other pages in Wikipedia. Alternating between the two taxonomies, edges are added to each taxonomy based on the information available in the other. Finally, heuristics further enrich the category taxonomy by adding hypernym edges for nodes which are still orphans after the first two steps. In contrast to both WikiTaxonomy and our work, WIBI ignores the syntactic structure of category titles.

3 Taxonomy induction

A unified, high-accuracy taxonomy of pages and categories is induced from WCN through the application of a cascade of linguistically motivated heuristics, which exploit lexical and structural information (mainly the lexical head of categories) from Wikipedia to generate a set of candidate generalizations for pages (*page heuristics*) and categories (*category heuristics*). Subsequently, other heuristics are used to simplify the taxonomy by eliminating redundant nodes (see Figure 1). The heuristics, which are derived empirically or adapted from previous work, are described in this section using these notations:

- E: set of all WCN edges;
- h_c: lexical head of the title of category c;
- $C_a(n)$: set of *all* direct parents of node n (page or category) in WCN, $\{c \mid (n,c) \in E\}$[2];
- $C_{pl}(n) \subset C_a(n)$: subset of parent categories ($C_a(n)$) whose titles have a *plural* lexical head, such as *Administrative divisions*. Categories with plural heads have played an important role in earlier work on taxonomy induction from Wikipedia, as they are more likely to be genuine classes (e.g. *Countries*) as opposed to instances (e.g. *France*) (Suchanek et al., 2007; de Melo and Weikum, 2010);

[2]Wikipedia maintenance categories (e.g., *Sports award stubs*) are removed using a handful of blacklisted keywords such as "articles", "stubs" etc.

• L_p: set of *defining* lemmas attached to the root copular verb in the first sentence of the Wikipedia description of page p, e.g., *"William Shakespeare was an English poet, ..."* (Flati et al., 2014);

• $sup(h_{c_1}, h_{c_2})$: *global support* for an ordered pair of lexical heads h_{c_1} and h_{c_2}, defined as the number of edges in E, from a category with head h_{c_1} to a category with head h_{c_2};

• $\overrightarrow{v_h}$: vector of co-occurrence counts of plural head h with every unique plural head h' in WCN, where co-occurrence count is defined as the number of pairs of categories with heads h and h' which have at least one common child (page or category);

• $tsim\,(h_1, h_2)$: type similarity, defined as the cosine similarity between $\overrightarrow{v_{h_1}}$ and $\overrightarrow{v_{h_2}}$;

3.1 Category heuristics

Same head. Similarly to the head-matching heuristic in Ponzetto and Strube (2007), for any category c, pick all categories $c' \in C_a(c)$ as candidate generalizations, if they have the same lexical head as c. E.g. *Category:American actors* is picked as candidate generalization for *Category:American child actors*.

Global head support. For any category c, pick the category $c' \in C_{pl}(c)$ with the highest[3] global support $sup(h_c, h_{c'})$ as a candidate generalization, provided the support is above a fixed threshold T_{sup}. E.g. *Category:American entertainers* is picked as candidate generalization for *Category:American actors* because $sup(\textit{actors}, \textit{entertainers}) > T_{sup}$.

Type similarity. For any category c, pick the category $c' \in C_{pl}(c)$ which has head h' with the highest[3] type similarity $tsim(h, h')$ as a candidate generalization, if the similarity is above a fixed threshold T_{tsim}. E.g. *Category:People by occupation* is picked as candidate generalization for *Category:Entertainers* because $tsim(\textit{entertainers}, \textit{people}) > T_{tsim}$.

Only plural parent. For any category c, if $C_{pl}(c)$ contains only one category, pick it as a candidate generalization.

Only singular parent. For any category c with a non-plural head h_c, if $C_a(c)$ contains only one category, pick it as a candidate generalization.

Grouping child category. Categories whose titles match the pattern **X by Y** (e.g. *"Actors by nationality"*) usually indicate groupings of instances of *class* X by *attribute* Y (Nastase and Strube, 2008). Thus, for category c whose title matches the pattern **X by Y**, pick the category with title **X** (if one exists) as a candidate generalization.

Grouping parent category. For any category c, pick those categories in $C_{pl}(c)$ as candidate generalizations, whose titles match the pattern **X by Y**. E.g. *Category:Occupations by type* is picked as candidate generalization for *Category:Legal professions*.

Suffix head. For any category c, pick all categories $c' \in C_{pl}(c)$, whose lexical heads $h_{c'}$ are suffixes of h_c, as candidate generalizations. E.g. *Category:People by occupation* is picked as candidate generalization for *Category:Sportspeople*.

Lookahead candidates. For any category c, pick its grandparents (second-level ancestor categories) as candidate generalizations, if they satisfy the conditions in the SAME HEAD, GROUPING PARENT CATEGORY or SUFFIX HEAD heuristics. Higher-level ancestors are ignored as they are usually inaccurate.

Title head. For any category c, pick the category with the title h_c as a candidate generalization, if the lemma of h_c is in top T_l% most frequent lemmas among the defining lemmas L_p of the child pages of c. E.g. *Category:Writers* is picked as candidate generalization for *Category:Legal Writers*[4].

[3] If multiple categories satisfy the condition, all of them are picked.

[4] Key difference between **Same head** and **Title head** heuristic is that the latter does not require the candidate generalizations to be present in $C_a(c)$.

3.2 Page heuristics

Exact defining lemma. For page p, pick the category $c \in C_{pl}(p)$ as a candidate generalization if the lemma of its lexical head is in L_p. E.g. all parent categories of page *Johnny Depp* with lexical head *actors* are picked as candidate generalizations because *actor* is present in $L_{\text{Johnny Depp}}$.

Type-similar lemma. For page p, pick a category $c \in C_{pl}(p)$ as a candidate generalization, if the type similarity between the category's lexical head (h_c) and at least one of the defining lemmas in L_p is greater than a fixed threshold T'_{tsim}. E.g. all parent categories of page *Johnny Depp* with lexical head *people* are picked as candidate generalizations because *actor* is present in $L_{\text{Johnny Depp}}$ and $tsim\,(actors, people) > T'_{tsim}$.

Plural head. Similar to YAGO (Suchanek et al., 2007), for page p, pick all categories in $C_{pl}(p)$ as candidate generalizations.

Transfer. If a page p has an equivalent category[5], pick candidate generalizations generated by category heuristics for the equivalent category as candidate generalizations of p.

3.3 Construction of the HEADS taxonomy

The heuristics[6] are applied to individual pages or categories in order of decreasing edge-level precision, as measured on a manually annotated development set, which is the same order in which they have been presented above. For each node, the process stops when one of the heuristics produces at least one generalization, and the remaining heuristics for that node are ignored. For example, in Figure 1a, only the generalizations proposed by α_p for entity *Tom Cruise* are retained, namely *Living people* and *Male actors from NY*. Certain categories encode information that is orthogonal to types, and therefore superfluous as it may refer to time (*20th-century actors*), location (*Actors from Singapore*) or grouping by attributes (*Actors by nationality*). Such categories are detected using a few regular expressions and eliminated: their children are linked directly to their parents, and the redundant nodes are removed (Fig. 1b) producing a more compact taxonomy (Fig. 1c). This step is hereafter referred to as *simplification*.

The described process results in the HEADS taxonomy, which is evaluated in the next section. Taxonomy generation and evaluation in this submission is restricted to English Wikipedia. However, it can be easily adapted to other languages by porting the heuristics, a fairly straightforward task if a dependency parser is available in the target language. Adaptation to other languages is not explored in this study, and remains the object of future work.

4 Taxonomy evaluation

This section compares the HEADS taxonomy against the state of the art. It presents the standard edge-level evaluation (Ponzetto and Strube, 2011; Flati et al., 2014); demonstrates that, as popular as they might be, edge-level metrics do not reflect the real quality of a taxonomy; and proposes a more comprehensive evaluation, which takes into account the correctness of multi-edge generalization paths, overall probability of generalization errors, granularity of individual generalizations and accuracy of specializations. It is shown that performance along these newly-proposed dimensions is not necessarily correlated with edge-level metrics and cannot be estimated directly from them.

Experimental setup HEADS is constructed using a November 2015 snapshot of the English Wikipedia. To create a baseline for comparison, we initially attempted to re-implement the state-of-the-art taxonomy induction approach of Flati et al. (2014), but were unable to replicate the reported results. In particular, recall of the re-implementation was lower than expected. Since the source code for WIBI was not made public and was not available upon request, we instead compared HEADS directly against the entity and

[5] A page and category are considered equivalent if they have the same title after lemmatization of each token. If a disambiguation string is specified in the title (e.g., *biology* in *Family (biology)*), it should also match. e.g., *Families (biology)* ~ *Family (biology)* ≁ FAMILY.

[6] Threshold T_{sup} is set to 5, T_{tsim} and T'_{tsim} are set to 0.2 and T_l is set to 10.

Taxonomy	WIBI$_E$	WIBI$_C$	HEADS
Nodes	3,414,512	597,179	4,580,662
Entities (E)	3,414,512	-	4,239,486
Categories (C)	-	597,179	341,176
Leaves	3,308,755	465,682	4,359,178
Edges	3,859,717	594,917	11,648,975
$E \rightarrow E$	3,859,717	-	-
$E \rightarrow C$	-	-	11,077,992
$C \rightarrow C$	-	594,917	570,983
Avg. degree	1.13	0.996	2.54
WCCs	6,448	2,301	3,195
	Largest WCC		
Nodes	3,386,995 (99.2%)	469,453 (78.6%)	4,563,949 (99.6%)
Edges	3,838,286 (99.4%)	469,453 (78.9%)	11,634,161 (99.9%)

Table 1: Topological properties of HEADS and WIBI taxonomies. (WCC: weakly connected component)

Taxonomy	Edge type	P	R	C	A
WCN	$E \rightarrow C$	0.785	1.000	1.000	0.902
	$C \rightarrow C$	0.807	1.000	0.970	0.840
HEADS	$E \rightarrow C$	0.394	0.249	0.898	0.956
	$C \rightarrow C$	0.405	0.344	0.249	0.931
WIBI$_E$	$E \rightarrow E$	0.841†	0.794†	0.926†	0.789
WIBI$_C$	$C \rightarrow C$	0.852†	0.829†	0.973†	0.840

Table 2: Edge-level evaluation. $E \rightarrow C$ represents entity→category edges, $E \rightarrow E$ represents entity→entity edges and $C \rightarrow C$ represents category→category edges. †: results as reported in Flati et al. (2014). P: precision, R: recall, C: coverage, A: accuracy.

category taxonomies made available by Flati et al. (2014), referred to as WIBI$_E$ and WIBI$_C$, respectively. It is important to stress that WIBI taxonomies are generated using an older Wikipedia snapshot (October 2012). However, to the best of our knowledge, there is no evidence suggesting that taxonomy induction is easier or harder on more recent vs. older snapshots. Noisy edges between categories such as *Japan↝660 BC* can be found in both snapshots. Meanwhile, the network has grown significantly, with more than twice as many categories (1.37M vs. 619K) and 20% more entities (4.7M vs 3.8M), possibly adding to the complexity of the task.

4.1 Topological properties

The main topological properties of the HEADS and WIBI taxonomies are shown in Table 1. HEADS contains fewer categories and category→category edges than WIBI$_C$, due to the simplification step (cf. Section 3.3), which removes approximately 53% of parent categories from WCN. HEADS covers a larger number of entities than WIBI taxonomies, but a direct comparison of absolute sizes is not necessarily meaningful since the three taxonomies are defined in different spaces (WIBI$_E$ has entity→entity edges, WIBI$_C$ has category→category edges, while HEADS has entity→category and category→category edges). In addition, as already mentioned, WIBI taxonomies are generated using an older snapshot of Wikipedia. As shown in Table 1, the largest weakly connected component in HEADS and WIBI$_E$ covers over 99% of the nodes. HEADS has 50% fewer components, which is desirable, as each component is an enclave of isolated entities. WIBI$_C$, which is an order of magnitude smaller than WIBI$_E$ and HEADS, has even fewer connected components, but is overall less connected, with the largest connected component containing only 78% of the nodes. Lastly, HEADS contains about twice as many edges per node as the WIBI taxonomies (see avg. degree), which allows it to better account for multiple aspects of a concept or an entity, e.g., *Johnny Depp* being both an *Actor* and a *Film producer*.

4.2 Edge-level evaluation

The first comparison between HEADS and WIBI taxonomies follows the methodology introduced and consistently followed in prior literature, namely computing edge-level precision and recall scores against a gold standard (Ponzetto and Strube, 2011; Flati et al., 2014). To build the gold standard, 500 entities and 500 categories are randomly selected, and their parents in WCN are annotated by three human judges as correct or incorrect generalizations. [7] Table 2 shows *precision* and *recall* scores for HEADS and WIBI taxonomies by edge type. Precision and recall with respect to the golden edges are computed for each sampled node, and then averaged over all the nodes in the gold standard.

Compared to the WIBI taxonomies, HEADS shows significantly lower precision and recall scores in this evaluation. However, the losses can be largely attributed to the simplification procedure (cf. Section 3.3). For example, in Figure 1, the edge *Tom Cruise→Male actors from NY* would be missing from the final HEADS taxonomy as the node *Male actors from NY* would be removed by the simplification

[7] The inter-annotator agreement in terms of Fleiss' Kappa is 0.52. Annotations were harmonized by majority voting.

procedure, thus resulting in loss of precision and recall. Similarly, *Living People→People*, a correct edge, would be considered a precision loss, as it is absent from WCN (and hence, from the gold standard).

Table 2 also reports *coverage*, defined as the fraction of entities and categories in a taxonomy with at least one generalization, independent of its correctness. HEADS shows lower coverage on categories, because 65% of categories in WCN are removed from HEADS due to the simplification procedure.

As an additional metric, Table 2 reports edge-level *accuracy*, defined as the ratio of edges annotated as correct over the total number of edges sampled from a taxonomy. Accuracy scores are computed for each taxonomy by randomly sampling 450 edges of each type and annotating their correctness. HEADS is more accurate than WIBI$_E$ for entities, though a direct comparison is not meaningful, as WIBI$_E$ contains entity→entity edges while HEADS contains entity→category edges. For category→category edges, HEADS achieves a fairly significant > 10% improvement in accuracy compared to WIBI$_C$ taxonomy.

4.3 Beyond edge-level evaluation

Good performance at edge level, though widely used as an indicator of quality for a taxonomy (Ponzetto and Strube, 2007; Nastase and Strube, 2008; Flati et al., 2014), does not automatically translate into good performance at path level. For example, the generalization path *apples→fruits⇝vegetarians→people→organisms* is 75% edge-accurate (i.e., 3/4 edges are correct as indicated by the symbol →), but it can lead to the wrong inference that *apples* are *vegetarians* and, in turn, *people* and *organisms*. A single incorrect edge, namely *fruits⇝vegetarians*, causes a cascade of generalization errors for *fruits* and all its descendants, and a cascade of specialization errors for *vegetarians* and all its ancestors.

As an alternative to edge-level evaluation, the remainder of this section proposes a more structured scheme for evaluating a taxonomy. More specifically, it seeks to estimate the following: (1) What is the accuracy of multi-edge generalization paths ? (2) Are individual generalizations at the right level of granularity ? and (3) What is the accuracy of specializations of a concept.

4.3.1 Path-level evaluation

From the above example (*apples⇝organisms*), it is clear that during traversal of an upward generalization path, the correctness of individual edges is inconsequential to finding a good generalization for starting node (i.e., *apples*) once the first wrong edge (*fruits⇝vegetarians*) is encountered. Therefore, a good taxonomy should not only provide a large proportion of correct edges, but also provide correct generalization paths, i.e., paths which are correct in their entirety. However, since in practice it is common for relatively deep taxonomies to provide long generalization paths which pick at least one wrong edge, it would be still desirable to have a long *correct path prefix*, i.e., the maximal prefix of a path which is correct in its entirety.

This section evaluates HEADS and WIBI taxonomies on their ability to provide longer correct path prefixes and correct generalization paths. To avoid bias, it is desirable that paths sampled from different taxonomies start from the same node. Therefore, WIBI$_C$, which lacks the notion of entities, is first augmented with $E{\rightarrow}C$ edges from HEADS, resulting in a new hybrid taxonomy hereafter referred to as WIBI$_C$+H$_E$. For a sample of 250 entities present in HEADS, WIBI$_E$ and WIBI$_C$+H$_E$, one upward path is sampled per entity per taxonomy, for a total of 750 paths. Example paths are shown in Table 3, while Figure 2 shows the length distribution of the generalization paths sampled from each taxonomy. As expected, HEADS paths are generally shorter than WIBI taxonomies due to simplification.

To compare the three taxonomies, three human annotators[8] inspect each path starting from the entity and annotate the first incorrect generalization (e.g., *Film producer⇝Filmmaking* for the WIBI$_E$ example in Table 3). Figure 3 shows the average length of the correct path prefix in HEADS and WIBI taxonomies, along with 95% confidence interval bars[9], depending on the total length of a path. For a correct generalization path, the length of correct path prefix is the same as the path length, so an ideal taxonomy

[8] At least two annotators agreed for 93% of paths. All three annotators agreed for 53% of paths. Annotations are harmonized using majority voting.

[9] The confidence intervals reflect the distribution of the paths being sampled. A larger confidence bar indicates lower probability that a path of that length is chosen.

WIBI$_E$	WIBI$_C$+H$_E$	HEADS
Structure	Government	
↑Algebraic structure	... 23 more categories ...	**Apes**
↑Category (mathematics)	↑Cinema by region	**↑Humans**
↑Sequence	↑Cinema by continent	**↑People**
↑Process (science)	↑North American cinema	**↑Producers**
↑Filmmaking	↑Cinema of the United States	**↑American producers**
↑Film producer	**↑American film producers**	**↑American film producers**
Johnny Depp	**Johnny Depp**	**Johnny Depp**

Table 3: Upward generalization paths for *Johnny Depp* in three taxonomies. Correct path prefixes are shown in bold.

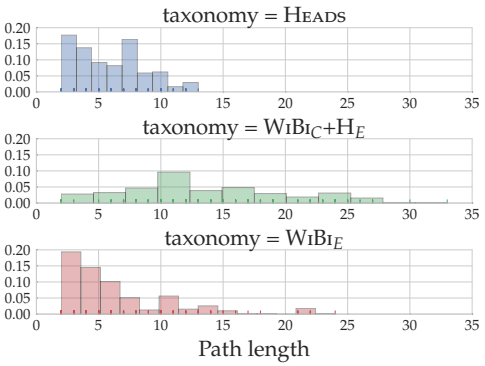

Figure 2: Length distribution of sampled generalization paths in different taxonomies.

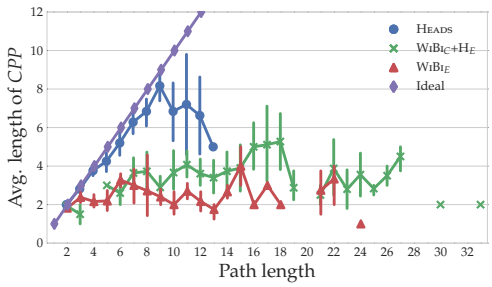

Figure 3: Average length of correct path prefix (*CPP*) in different taxonomies (computed using 750 annotated paths).

with only correct generalization paths would show up as the line $y = x$ in Figure 3. The behavior of HEADS is very close to an ideal taxonomy for the majority of path lengths, and outperforms WIBI$_E$ or WIBI$_C$+H$_E$ at all lengths, while WIBI$_C$+H$_E$ slightly outperforms WIBI$_E$. It is interesting to note that this difference does not translate into similar differences in edge-level evaluation, where all taxonomies consistently show relatively high accuracy (cf. Section 4.2). The superior performance of HEADS is further confirmed by Figure 4, which shows the probability of obtaining a correct generalization path of length $\leq k$. In contrast with WIBI taxonomies, HEADS generalization paths maintain high probability of correctness (> 0.7) at all lengths.

4.3.2 Path-granularity evaluation

A good taxonomy should not only provide correct generalization paths, but also ensure that each individual edge in the path provides generalization at the right level of granularity, i.e., neither too specific nor too general. To evaluate this aspect, 100 generalization paths originating from the same starting entities are sampled from different taxonomies. For each path, each individual edge is annotated by three human annotators with one of the following labels: 0 for wrong generalization (*fruits⇝vegetarians*); 1 for under-generalization (*fruits by country→fruits*); 2 for good-generalization (*edible fruits→fruits*); 3 for over-generalization (*edible fruits→physical bodies*). An edge under-generalizes if it adds or removes little information relative to the source node (e.g. *cricketers by team→cricketers*) or if it is a synonym or rephrasing of the original category (e.g. *coaches by sport→sport coaches*). An edge over-generalizes if it removes too much information. For example, for *bitstream→concept* one would expect the taxonomy to provide additional intermediate nodes (e.g. *binary sequences*) before generic node *concept*. Good-generalization label implies that edge is correct and neither over-generalizes nor under-generalizes. In order to ensure that the paths on which the comparison is performed are similar in length and complexity, we only consider pairs of shortest paths $\langle p_1, p_2 \rangle$ with the same final node, selected so as to minimize the difference in the length of the shortest paths $||p_1| - |p_2||$ in the two taxonomies, while ensuring that the paths are not identical ($p_1 \neq p_2$).

WIBI$_E$ is excluded from this experiment, since in contrast to HEADS and WIBI$_C$+H$_E$, WIBI$_E$ does not contain categories, hence the condition of same final node cannot be satisfied. Figure 5 graphically summarizes the results of this experiment. HEADS has fewer under-generalizations than WIBI$_C$+H$_E$

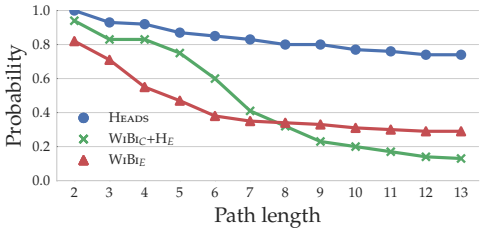

Figure 4: Probability of correct generalization paths vs. length (computed using 750 annotated paths). The probability at length k is the ratio of correct paths of length $\leq k$ to the total number of paths of length $\leq k$. Paths with length > 13 are omitted, as they are not present in HEADS samples, and always incorrect in WIBI$_E$ and WIBI$_C$+H$_E$ samples.

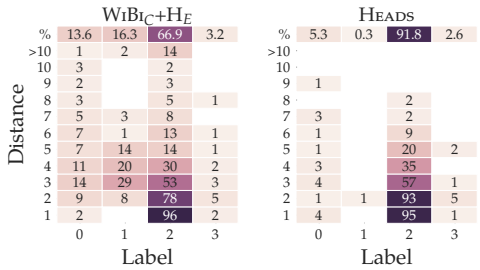

Figure 5: Generalization granularity evaluation for HEADS and WIBI$_C$+H$_E$ using 100 generalization paths. Labels on the horizontal axis indicate generalization granularity: 0 (wrong generalization), 1 (under-generalization), 2 (good-generalization) and 3 (over-generalization). Top row shows overall distribution of labels. Other rows represent number of sampled paths which have an edge with the corresponding label at the given distance from starting node.

Taxonomy	Overall accuracy	Per-node accuracy
WIBI$_E$	0.243	0.230
HEADS (entity)	**0.703**	**0.727**
WIBI$_C$	0.381	0.408
HEADS (category)	**0.670**	**0.725**

Table 4: Accuracy of specializations (computed using 100 (node, descendant) pairs). Overall accuracy is fraction of sampled (node, descendants) pairs which are correct, and per-node accuracy represents the average ratio of correct descendants per node. Results for entity and category descendants of HEADS are reported separately.

(0.3% vs 16.3%), which can be largely attributed to the simplification procedure (cf. Section 3.3). Despite the removal of 65% of categories through simplification, HEADS still does not suffer significantly from over-generalizations.

4.3.3 Specializations evaluation

A good taxonomy provides not only accurate generalizations going upwards in the taxonomy, but also accurate specializations going downwards. To evaluate this aspect, three human annotators annotate the correctness of a sample of descendants, for nodes in the taxonomies WIBI$_E$, WIBI$_C$ and HEADS. To avoid bias, nodes (entities for WIBI$_E$; categories for WIBI$_C$, HEADS) are sorted in decreasing order of the number of descendants in the respective taxonomies. 10 nodes at fixed ranks (5, 10, .., 50) from each list are selected for evaluation. To enable a comparison of WIBI$_E$ with WIBI$_C$ and HEADS, category nodes are manually mapped to equivalent entity nodes and vice-versa (e.g., *Category:Concepts* is mapped to the entity *Concept*). The annotators judge the correctness of 10 randomly sampled descendants for each selected node in each of the three taxonomies (see Table 4). HEADS is almost three times as accurate for entities as WIBI$_E$, and almost twice as accurate for categories as WIBI$_C$.

4.3.4 Extrinsic evaluation

This section compares HEADS, WIBI$_E$ and WIBI$_C$+H$_E$ on the task of selecting correct generalizations (e.g., *Countries*) for the variable slot in lexicalized templates such as *Passport of [X]*. These templates are mined by aggregation of Wikipedia page titles (e.g., *Passport of France*, *Passport of Canada*). The lexical fillers observed in the titles (e.g., *France*, *Canada*) are automatically disambiguated to a specific page (e.g., *France, European country* rather than *France, NY town*)[10], resulting in a set of filler entities for a template referred to as the template *support*.

To evaluate a taxonomy, for each template, the taxonomy is repeatedly traversed starting from subsamples of the support entities and equal-sized samples of random entities. Each non-leaf node in the taxonomy receives a score equal to the difference between the counts of support entities versus random

[10]Details of non-trivial problems of template mining and filler disambiguation are omitted due to space constraints, as they are not the main focus of this paper.

Template	Selected Generalizations		
	WIBI$_E$	WIBI$_C$+H$_E$	HEADS
railways in [X]	Tool, Entity, Publication, Operation (mathematics), Property (philosophy), Administrative division, Fine art, Material, Wealth, Combination	Geography, Countries, Statistics, Mathematical and quantitative methods (economics), Least developed countries, Capitals	Cities, Least developed countries, Administrative territorial entities
forestry in [X]	Entity, Wealth	Muslim-majority countries, Geography, Countries, Statistics, Mathematical and quantitative methods (economics), French-speaking countries and territories, Least developed countries	Muslim-majority countries, Least developed countries, Administrative territorial entities
[X] reader	Economic system, Entity, Document, Property (philosophy), Fine art, Material, Wealth	Philosophical concepts, Branches of philosophy, Concepts in metaphysics, Digital technology, Society, Psychology, Intelligence, Classification systems	Intellectual works, Concepts, Storage media, Literary characters
[X] ' day	Plurality (voting)	Public economics, Heavy metal subgenres, Intelligence, Economic policy, Heavy metal musical groups by nationality	Social groups, Occupations, Creative works
tomb of [X]	Tool, Value (mathematics), Publication, Proclamation, Official, Document, Instance (computer science), Fine art, Capital (economics), Material, Aesthetics, Electoral district, *[+2 more]*	People by nationality, Countries by continent, Hebrew Bible people, Ancient people, Religion, Genetics, Behavior, People by occupation, Fields of application of statistics, Jewish priests, Monarchy, Statistics, *[+16 more]*	People, Families, Ethnic groups, Noble titles

Table 5: Lists of selected generalizations for HEADS and WIBI taxonomies.

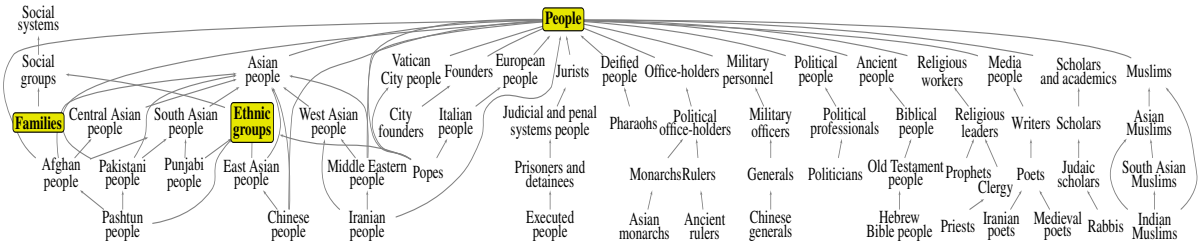

Figure 6: A view of the largest connected component of HEADS explored during the generalization of the template *Tomb of [X]*. The highlighted nodes are the selected generalizations.

entities from which the node can be reached in the given taxonomy. A node is selected as the generalization of fillers for the template, if its score is higher than both 1) the score of any of its parents, and 2) the sum of the scores of its children. For example, Figure 6 shows a subset of the largest connected component of HEADS explored while generalizing the fillers of the template *Tomb of [X]*. The selected generalizations are highlighted. The rationale behind this process is that in a good taxonomy, entities in the support (e.g., *France*, *Canada*) should consistently activate the same set of good generalizations (e.g., *Countries*).

Table 5 shows the lists of generalizations obtained with WIBI$_E$, WIBI$_C$+H$_E$ and HEADS for a few templates. A quantitative comparison of the results is inherently complex and outside the scope of this paper, yet it is immediately apparent that the generalizations obtained with WIBI$_C$+H$_E$ and WIBI$_E$ are generally quite noisy (e.g., *Genetics* for *Tomb of [X]*, *Wealth* for *railways in [X]*). On the contrary, HEADS shows a superior ability to select meaningful and compact generalizations that account for the polysemy of the templates without sacrificing precision (e.g., *People*, *Families*, *Ethnic groups* and *Noble titles* for *Tomb of [X]*).

5 Conclusion

Whether built from scratch or derived by filtering existing data, automatically-constructed taxonomies are accurate and useful only to the extent that they correctly assert not only short-range, but also longer-range generalizations among concepts or entities. The unified taxonomy introduced in this paper assembles entities and categories from Wikipedia that are in *is-a* relation relative to one another, primarily by detecting and analyzing lexical heads. A thorough evaluation framework is presented, and applied to the new taxonomy. In every respect, the taxonomy represents a significant improvement over the state of the art. It is more accurate along paths of arbitrary length and provides more accurate specializations.

References

G. de Melo and G. Weikum. 2010. MENTA: Inducing multilingual taxonomies from Wikipedia. In *Proceedings of the 19th International Conference on Information and Knowledge Management*, pages 1099–1108.

X. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, and K. Murphy. 2014. Knowledge Vault: A Web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 601–610, New York, NY.

David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A Kalyanpur, Adam Lally, J William Murdock, Eric Nyberg, John Prager, et al. 2010. Building Watson: An overview of the DeepQA project. *AI Magazine*, 31(3):59–79.

T. Flati, D. Vannella, T. Pasini, and R. Navigli. 2014. Two is bigger (and better) than one: the Wikipedia Bitaxonomy project. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 945–955, Baltimore, Maryland. Association for Computational Linguistics.

M. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th Conference on Computational Linguistics*, pages 539–545, Nantes, France.

J. Hoffart, F. Suchanek, K. Berberich, and G. Weikum. 2013. YAGO2: a spatially and temporally enhanced knowledge base from Wikipedia. *Artificial Intelligence Journal. Special Issue on Artificial Intelligence, Wikipedia and Semi-Structured Resources*, 194:28–61.

E. Hovy, R. Navigli, and S. Ponzetto. 2013. Collaboratively built semi-structured content and Artificial Intelligence: The story so far. *Artificial Intelligence*, 194:2–27.

J. Hu, G. Wang, F. Lochovsky, J. Sun, and Z. Chen. 2009. Understanding user's query intent with Wikipedia. In *Proceedings of the 18th World Wide Web Conference*, pages 471–480, Madrid, Spain.

J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. Mendes, S. Hellmann, M. Morsey, P. van Kleef, and S. Auer et al. 2015. DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, 6(2):167–195.

R. Mihalcea. 2007. Using Wikipedia for automatic word sense disambiguation. In *Proceedings of the 2007 Conference of the North American Association for Computational Linguistics*, pages 196–203, Rochester, New York.

V. Nastase and M. Strube. 2008. Decoding Wikipedia categories for knowledge acquisition. In *AAAI*, volume 8, pages 1219–1224.

V. Nastase and M. Strube. 2013. Transforming Wikipedia into a large scale multilingual concept network. *Artificial Intelligence*, 194:62–85.

V. Nastase, M. Strube, B. Börschinger, C. Zirn, and A. Elghafari. 2010. WikiNet: A very large scale multi-lingual concept network. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, pages 17–23, La Valetta, Malta.

S. Ponzetto and M. Strube. 2007. Deriving a large scale taxonomy from Wikipedia. In *Proceedings of the 22nd National Conference on Artificial Intelligence*, pages 1440–1445, Vancouver, British Columbia.

S. Ponzetto and M. Strube. 2011. Taxonomy induction based on a collaboratively built knowledge repository. *Artificial Intelligence*, 175(9–10):1737–1756.

L. Ratinov and D. Roth. 2012. Learning-based multi-sieve co-reference resolution with knowledge. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1234–1244, Jeju Island, Korea.

L. Ratinov, D. Roth, D. Downey, and M. Anderson. 2011. Local and global algorithms for disambiguation to Wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 1375–1384, Portland, Oregon.

M. Remy. 2002. Wikipedia: The free encyclopedia. *Online Information Review*, 26(6):434.

F. Suchanek, G. Kasneci, and G. Weikum. 2007. Yago: A core of semantic knowledge unifying Wordnet and Wikipedia. In *Proceedings of the 16th International World Wide Web Conference*, pages 697–706, Banff, Canada.

D. Vrandečić and M. Krötzsch. 2014. Wikidata: A free collaborative knowledge base. *Communications of the ACM*, 57:78–85.

F. Wu and D. Weld. 2010. Open information extraction using Wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 118–127, Uppsala, Sweden.