Evaluation Dataset and Methodology for Extracting Application-Specific Taxonomies from the Wikipedia Knowledge Graph

Georgeta Bordea, Stefano Faralli, Fleur Mougin, Paul Buitelaar, Gayo Diallo

Evaluation Dataset and Methodology for Extracting Application-Specific Taxonomies from the Wikipedia Knowledge Graph

Georgeta Bordea[1], Stefano Faralli[2], Fleur Mougin[1], Paul Buitelaar[3], and Gayo Diallo[1]
[1] University of Bordeaux, Inserm, Bordeaux Population Health Research Center, ERIAS team, France
[2]University of Rome Unitelma Sapienza, Italy
[3]Data Science Institute, NUI Galway, Ireland
{georgeta.bordea, fleur.mougin, gayo.diallo}@u-bordeaux.fr,
stefano.faralli@unitelmasapienza.it, paul.buitelaar@nuigalway.ie

Abstract

In this work, we address the task of extracting application-specific taxonomies from the category hierarchy of Wikipedia. Previous work on pruning the Wikipedia knowledge graph relied on silver standard taxonomies which can only be automatically extracted for a small subset of domains rooted in relatively focused nodes, placed at an intermediate level in the knowledge graphs. In this work, we propose an iterative methodology to extract an application-specific gold standard dataset from a knowledge graph and an evaluation framework to comparatively assess the quality of noisy automatically extracted taxonomies. We employ an existing state-of-the-art algorithm in an iterative manner and we propose several sampling strategies to reduce the amount of manual work needed for evaluation. A first gold standard dataset is released to the research community for this task along with a companion evaluation framework. This dataset addresses a real-world application from the medical domain, namely the extraction of food-drug and herb-drug interactions.

Keywords: Taxonomy Extraction, Knowledge Graphs, Knowledge Graph Pruning

1. Introduction

Taxonomies are useful tools for content organisation, navigation, and information retrieval, providing valuable input for semantically intensive tasks such as question answering (Ojokoh and Adebisi, 2019) and textual entailment (Fawei et al., 2019). Recent efforts on setting common grounds for taxonomy extraction evaluation in the context of SemEval have focused on extracting hierarchical relations from unstructured text (Bordea et al., 2015; Bordea et al., 2016; Jurgens and Pilehvar, 2016). In the past few years, this task attracted an increasing amount of interest from the research community. But the relatively low evaluation results show that this is still a challenging task, despite efforts to simplify the task by focusing on the more simple subtask of hypernym relations extraction rather than constructing a full fledged taxonomy and by providing datasets that mainly cover concepts with a relatively high frequency in the target corpus (Camacho-Collados et al., 2018). Instead of extracting hierarchical relations from text alone, in this work, we address the problem of adapting and re-using existing taxonomical structures from general-purpose knowledge graphs such as the Wikipedia knowledge graph (Kapanipathi et al., 2014).

The sheer vastness of Wikipedia's domain coverage recommends it as a reliable source of semi-structured information. Manually-curated and high-quality information about hierarchical relations is readily available for a wide range of domains. Hence, the taxonomy extraction task should be made more feasible provided that there are methods to accurately identify relevant information for a given domain or application scenario. The challenge is to deal with noisy relations when selecting appropriate concepts and relations because the Wikipedia category hierarchy makes use of loosely defined relations between an article and a category, easily leading to semantic drift. In this work, we move away from the tradition of evaluating extracted taxonomies on a few broad and unrelated domains, such as music, finance, or sports and we focus instead on a real-world application scenario from the medical domain: the discovery of food-drug interactions and herb-drug interactions from scientific articles. Foods and medicinal plants, for example *grapefruit juice* and *St. John's Wort*, potentially cause clinically-significant drug interactions in a similar way that combining drugs can lead to undesired drug-drug interactions. In this context, the goal is to automatically extract application-specific taxonomies from the Wikipedia category hierarchy for concepts that describe foods, plants and drugs in relation to diseases, signs, symptoms and medical specialties. Identifying these hierarchies is useful for constructing relevant corpora (Bordea et al., 2018; Bordea et al., 2019), entity extraction, and for drug interaction organisation and visualisation. But selecting a taxonomy of relevant Wikipedia categories is not trivial for this application because many of these concepts have several uses outside of our application domain. For instance, exploring *Category:Foods*, we discover *Category:Ducks* that points to *Category:Fictional Ducks*, including *Donald Duck* with all the films, television series and video games that are related, but clearly out of scope for our application. Similarly, *Category:Blood_sausages* is linked to *Category:Vampirism* through the category *Category:Blood_as_food*.

Previous work on pruning knowledge graphs (Faralli et al., 2018b) relies, for a comparative evaluation, on taxonomies extracted for a few carefully handpicked domains such as *Category:Singers*, *Category:Entertainers* and *Category:Companies*. These subgraphs are extracted through a basic approach based on the transitive closure of the root node and can be directly used as silver-standard datasets. But this approach is not guaranteed to produce an error-free, gold-standard category hierarchy, even for very fo-

cused concepts such as the ones selected by Faralli et al. Instead, we introduce a first gold standard dataset that covers three related domains in the context of a realistic application.
The key contributions of our work are as follows:

1. An iterative methodology for manually identifying domain-specific categories using a state-of-the-art algorithm for extracting taxonomies from Wikipedia.

2. A set of application-specific gold standard taxonomies produced with the above extraction methodology.

3. An evaluation framework to compare against gold standard datasets extracted from the Wikipedia knowledge graph.

This paper is organised as follows: first, we present our taxonomy extraction approach that extends a state-of-the-art algorithm for extracting application-specific taxonomies in section 2. Then, we present an iterative extraction methodology that is required to progressively filter irrelevant nodes as described in section 3. In section 4., we present a companion evaluation framework for the task of extracting taxonomies from the Wikipedia category structure and finally, we conclude this work in section 5.

2. Taxonomy extraction approach

In this section, we describe the proposed approach for the extraction of application-specific taxonomies from the Wikipedia knowledge graph.
Our approach is composed of two main steps (Figure 1):

1. **Initial domain/application specific selection.** The aims of this step are to: i) select a domain/application specific set of leaf nodes, $leaves$, to be the ground of the desired taxonomy; ii) select a proper node R to be the root node of the final taxonomy; iii) create a directed graph G including the relations between Wikipedia categories and sub-categories and between categories and Wikipedia pages (Figure 2).

2. **Iterative extraction.** When both the initial leaf nodes $leaves$ and the root node R have been selected, we start an iterative procedure where, during each iteration i: i) we apply the pruning algorithm described below to the initial graph G obtaining a Directed Acyclic Graph (DAG) G_i, and ii) we evaluate the domain pertinence of a sample of intermediate nodes of G_i, and (if needed) we define two sets, namely $keep$ and $discard$, containing nodes that the pruning algorithm will respectively preserve and discard at the next iteration. The size of the extracted graph is prohibitively large for an exhaustive manual verification, therefore we propose several strategies for sampling nodes based on their importance in the graph (see section 2.2.). The end of the iterative phase is determined when the size of the graph allows a complete manual assessment of the extracted nodes and the graph contains only domain/application specific nodes.

2.1. CrumbTrail

CrumbTrail (Faralli et al., 2018a; Faralli et al., 2018b) is an efficient and highly accurate algorithm to prune noisy or over-ambiguous knowledge graphs given as input an extensional definition of a domain of interest, namely as a set of instances or concepts. The algorithm climbs the graph in a bottom-up fashion, iteratively layering the graph and pruning nodes and edges in each layer while not compromising the connectivity of two sets of input nodes: i) the set of $leaves$ nodes representing the ground of the targeted application domain; ii) the set of $keep$ nodes, namely intermediate nodes that have not been removed while pruning and that will further be considered as preferential nodes to climb on in the next iteration. When applied to broad domains, the CrumbTrail algorithm produces a large number of out-of-domain nodes. We modified the original implementation of the algorithm to allow as input another set of nodes, $discard$. These nodes are simply removed from the graph before starting the pruning. In Figures 3 and 4, we depict an example of a noisy graph and the resulting DAG provided by CrumbTrail, respectively.

2.2. Iterative node filtering

The CrumbTrail algorithm is extracting a large number of categories that are not relevant to the considered domains. Further manual filtering is required, but checking every category manually is not feasible due to the large size of the graph. In general, we consider a node to be relevant if there is a straightforward hypernymy relation with the root of the taxonomy (e.g., *antibiotics* are *drugs*), but we allow as well broader categories such as *Health* or *Endocrinology*. Although the relation is not as direct, we can assume that the *Endocrinology* category is useful to classify drugs used in this branch of medicine. Similarly, we consider different cuisines based on geography, such as *Japanese cuisine*, to be relevant for the FOODS domain. To address these particularities, we employ several strategies to progressively discard out-of-domain nodes, with the goal of reducing as much as possible the number of categories that have to be manually analysed at each step. The following strategies are proposed to achieve this:

- **Top level sampling.** Identify irrelevant nodes from the first two levels of categories starting from the *Main_topic_classifications* Wikipedia node[1]. Wikipedia uses this category to group major topic classifications in one place and as a main entry point to the category hierarchy. The advantage is that a relatively small number of categories have to be analysed at the top levels of the hierarchy.

- **Centrality sampling.** Top ranked nodes based on Eigen vector centrality calculated with 100 iterations are manually analysed and removed if not relevant. Centrality-based methods have been previously used to identify topic labels from Wikipedia (Hulpus et al., 2013).

- **String-based sampling.** Nodes are discarded if they contain one of the following strings: *area*, *asia*, *chris-*

[1] https://en.wikipedia.org/wiki/Category:Main_topic_classifications

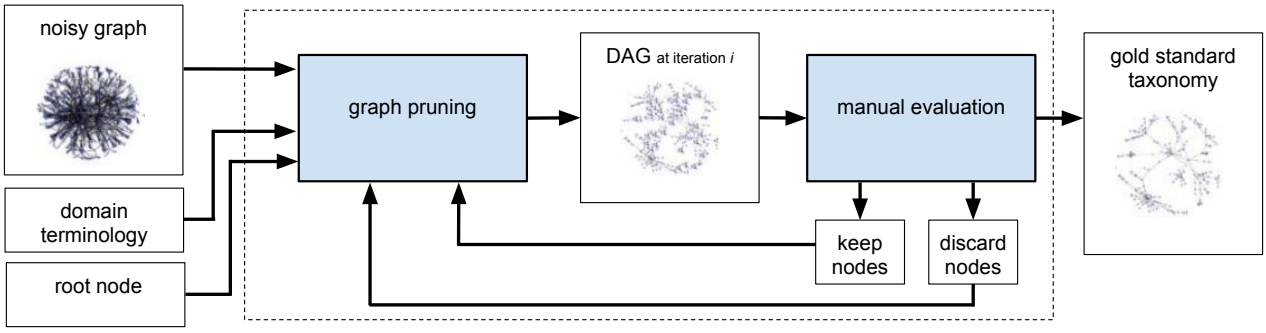

Figure 1: Workflow of the proposed iterative methodology for the extraction of application-specific taxonomies from the Wikipedia knowledge graph.

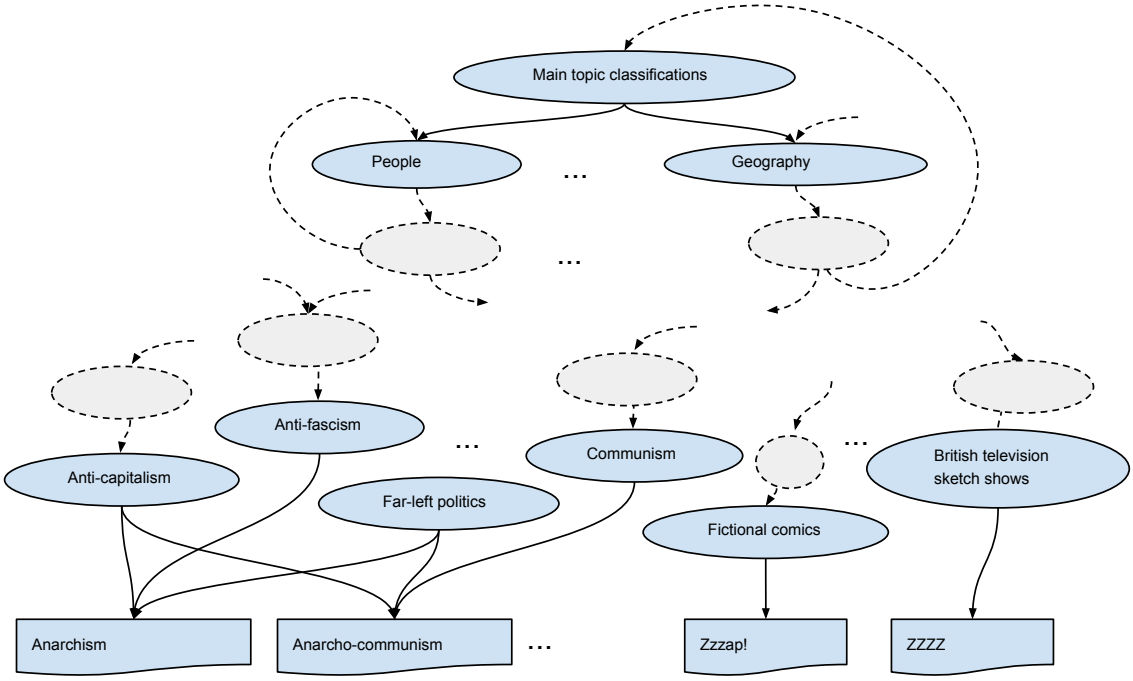

Figure 2: Structure of the initial graph built on top of Wikipedia pages and categories. The graph contains cycles and, in general, its complexity does not enable domain/application experts to extract manually a taxonomy of interest.

tian, *cities*, *companies*, *computer*, *countr*, *economy*, *educa*, *europe*, *film*, *geography*, *history*, *lakes*, *law*, *list*, *media*, *music*, *parties*, *people*, *philosopher*, *place*, *political*, *religi*, *software*, *sport*, *states*, *television*, *united*, *universities*, *war*, *writer*, *years*. Additionally, we remove nodes that start with a cardinal number and nodes that end with the string *_culture*. The list has been empirically populated by identifying from a random sample of 3,000 Wikipedia categories substrings frequently associated with unrelated domains.

- **Exhaustive filtering.** The complete list of nodes is analysed to identify irrelevant nodes, when feasible based on the size of the graph. Several iterations might be required because the additional nodes are discovered based on the discarded nodes.
- **Border community sampling.** A community analysis of the graph based on modularity shows that the majority of relevant nodes are densely connected in one modularity class. We define border community as a group of nodes that are placed near the border of the main relevant community. The intuition is that nodes that are further away in the graph are less likely to be pivotal in the creation of the gold standard.

The first three sampling approaches are used for filtering noisy nodes on successively smaller graphs, while the last approach is used as a recovery step and is performed on the initial noisy graph. In our experiments, we make use of a tool to visually analyse the nodes in the graph (Bastian et al., 2009). In general, this methodology is not domain-specific and can be directly applied to other specific domains, with the exception of the string-based sampling step. The list of strings proposed above is specific to our application domain and should be revised and extended accordingly for other domains.

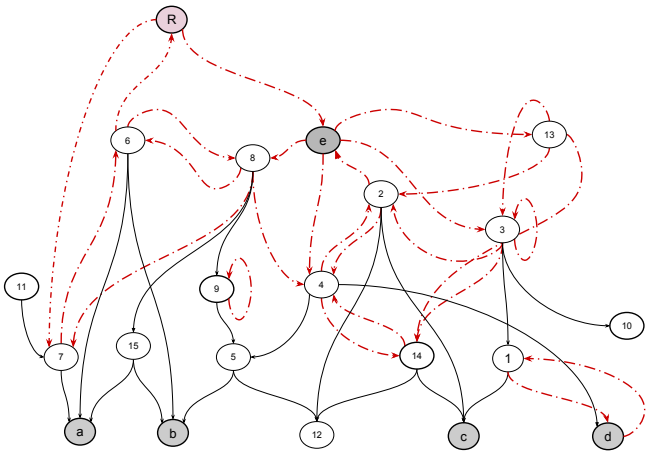

Figure 3: An example of a noisy graph from the paper (Faralli et al., 2018a) where dashed edges are part of cycles, R is the selected root node, $keep = \{e\}$, and $leaves = \{a, b, c, d\}$.

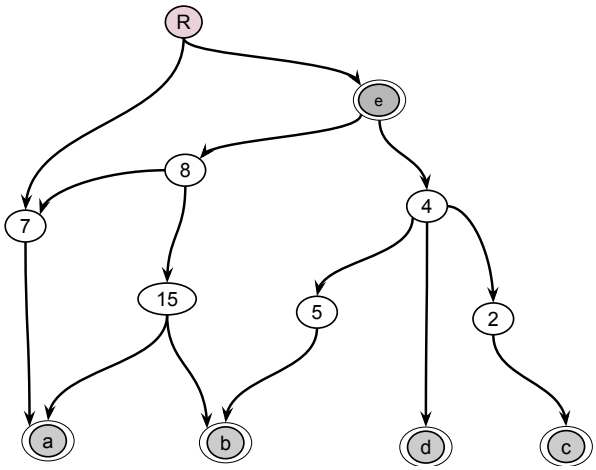

Figure 4: An example of a pruned graph from the paper (Faralli et al., 2018a), R is the selected root node, $keep = \{e\}$, and $leaves = \{a, b, c, d\}$.

3. Gold standard creation

In this section, we describe the domain/application specific taxonomies extracted from the Wikipedia knowledge graph with the methodology introduced in section 2. We first describe the input articles for each domain and then we present how we iteratively use CrumbTrail to extract gold standard taxonomies from Wikipedia.

3.1. Domain terminologies selection

Our resource includes three domain taxonomies: DRUGS, PLANTS and FOODS. The three initial domain terminologies provided as input for iterative extraction are manually collected from two reference compendia about interactions of drugs with foods and herbal medicines, respectively (Baxter and Preston, 2010; Baxter et al., 2013). Both compendia provide an index of interactions mentioning a drug and the food or plant with which they are potentially interacting. Next, we manually find the Wikipedia article that describes each item to be further provided as input for the CrumbTrail algorithm. In this way, we identify 215 Wikipedia articles about drugs, 70 Wikipedia articles about foods, and 134 articles about plants that are either known or suspected to be involved in drug interactions. The roots selected for these domains are the Wikipedia categories *Drugs*[2], *Foods*[3], *Plants*[4], and *Pharmacology*[5]. We also considered using the *Medicinal_plants* category as root node for the *Plants* domain but the resulting hierarchy was too small for our purposes.

3.2. Iterative filtering

We performed a total of 8 filtering iterations for *DRUGS*, 7 filtering iterations for *FOODS*, and 8 filtering iterations for *PLANTS*. Typically, the first filtering iteration makes use of the top level sampling strategy. For as long as the resulting list of nodes is too large to be manually evaluated in its entirety, we apply the centrality sampling and the string-based approach. The last filtering iteration is based on an exhaustive analysis of nodes. Table 1 gives an overview of the number of nodes in the *keep* or *discard* sets in each iteration and the size of the taxonomy in terms of nodes before and after the filtering. Based on our resources for manual analysis, a list of nodes is considered too large to be exhaustively analysed if there are more than 2,500 nodes in the taxonomy. We estimate that over the first three iterations where the CrumbTrail algorithm is the most productive, the amount of manual work required to evaluate the taxonomies is reduced to one percent when using the proposed filtering approach.

Our approach is more effective for removing noisy nodes than for identifying preferred categories, with only a small list of nodes identified for the *keep* set. For example, in the first filtering iteration, we identified the top level category *Health* for the DRUGS domain, and the *Nature* and *Health* categories for the FOODS and PLANTS domains. In the last filtering iteration, we identified the *Abortion* and *Midwifery* categories as potentially interesting for the PLANTS domain as they are related to possible uses of medicinal plants.

Finally, we performed the recovery step using the border community sampling approach. We manually evaluate 309 noisy nodes for the *DRUGS* domain, 598 noisy nodes for the *FOODS* domain and 763 noisy nodes for the *PLANTS* domain. This allows us to discover 33 *keep* nodes, 90 *keep* nodes, and 85 *keep* nodes for the three domains, respectively. Table 2 gives several statistics about the size of the final gold standard taxonomies for the three domains. The largest gold standard taxonomy was obtained for the *DRUGS* domain, followed by the gold standard taxonomy for the *PLANTS* domain.

Figure 5 gives an overview of the extracted gold standard taxonomies, showing the top level categories and the overall structure of the gold standards. Compared to the initial noisy graph, a relatively small number of nodes survived the iterative pruning process. Categories related to geographical locations were generally not considered to be relevant for this application, for example

[2]https://en.wikipedia.org/wiki/Category:Drugs
[3]https://en.wikipedia.org/wiki/Category:Foods
[4]https://en.wikipedia.org/wiki/Category:Plants
[5]https://en.wikipedia.org/wiki/Category:Pharmacology

Domain	Iteration	Filtering	Keep nodes	Discard nodes	Input	Output
Drugs	1	Top level	1	186	19977	5,625
	2	Centrality	0	20	5,625	2,241
	3	Centrality + String-based	0	33	2,241	1,528
	4	Centrality + String-based	0	79	1,528	1,005
	5	Exhaustive + String-based	0	213	1,005	2,386
	6	Exhaustive + String-based	0	356	2,386	1,666
	7	Exhaustive + String-based	0	624	1,666	1,380
	8	Exhaustive + String-based	0	234	1,380	805
	9	Recovery	33	0	805	838
Foods	1	Top level	2	34	17141	921
	2	Centrality	0	20	921	10,141
	3	Centrality + String-based	0	48	10,141	24,181
	4	Centrality + String-based	0	161	24,181	6,285
	5	Centrality + String-based	0	286	6,285	2,073
	6	Exhaustive + String-based	0	1,434	2,073	1,010
	7	Exhaustive + String-based	0	451	1,010	331
	8	Recovery	90	0	331	421
Plants	1	Top level	2	23	20119	11,902
	2	Centrality	0	20	11,902	899
	3	Centrality + String-based	0	11	899	784
	4	Centrality + String-based	0	68	784	554
	5	Exhaustive + String-based	0	60	554	837
	6	Exhaustive + String-based	0	343	837	1,454
	7	Exhaustive + String-based	0	745	1,454	1,549
	8	Exhaustive + String-based	2	884	1,549	443
	9	Recovery	85	0	443	528

Table 1: Size of discard, keep, input and output sets of nodes for different filtering iterations by domain.

Biota_of_[geographical location], *Flora_of_[geographical location]* or *Health_in_[geographical location]*. On the contrary, categories that follow the pattern *[geographical location]_cuisine* were considered to be relevant for the *FOODS* domain.

Domain	Nodes	Edges	Diameter	Avg. path
Drugs	805	1,674	16	4.369
Foods	331	690	8	3.001
Plants	443	768	12	4.312

Table 2: Gold standard taxonomies statistics including number of nodes, number of edges, diameter of the graph and average path length.

	Fleiss k		Precision	
Domain	Removed	Survived	Removed	Survived
Drugs	0.99	0.96	0.986	0.987
Foods	0.99	0.90	0.995	0.966
Plants	0.99	0.90	0.996	0.935

Table 3: Inter-annotator agreement *Fleiss k* on the six annotated datasets for the assessment of node domain pertinence of removed and survived nodes, performed by three domain experts.

3.3. Gold standard quality evaluation

For each resulting taxonomy, we performed a manual assessment of the correctness of a sample of nodes. We asked three human annotators to evaluate the domain relevance of a sample of 300 nodes that are removed after the first iteration and 300 nodes that survived the iterative pruning to be included in the gold standard (see section 2.). In Table 3, we report the corresponding agreement (Fleiss k) between the three annotators when evaluating relevance for the removed and survived nodes. In the same table, the last two columns show the average precision regarding correctly removed (not pertinent) nodes and correctly not removed (pertinent) nodes. The agreement between annotators is generally high, with values close to the upper range of almost perfect agreement. This shows that for humans this is a relatively easy task that can be further automated, but in general the agreement is higher for removed nodes than for survived nodes, indicating higher subjectivity.

4. Benchmark evaluation framework

In this section, we describe the measures we propose to compare a taxonomy against our three gold/silver domain taxonomies.

Given an automatically induced taxonomy $I = (V_I, E_I)$ and a gold/silver standard taxonomy $G = (V_G, E_G)$, the evaluation frameworks includes:

- a measure to assess the edge precision ($P_E = \frac{|E_G \cap E_I|}{|E_I|}$);
- a measure to assess edge recall ($R_E = \frac{|E_G \cap E_I|}{|E_G|}$);
- the *Cumulative F&M* measure (Velardi et al., 2012) which traverses the gold standard taxonomy G and the

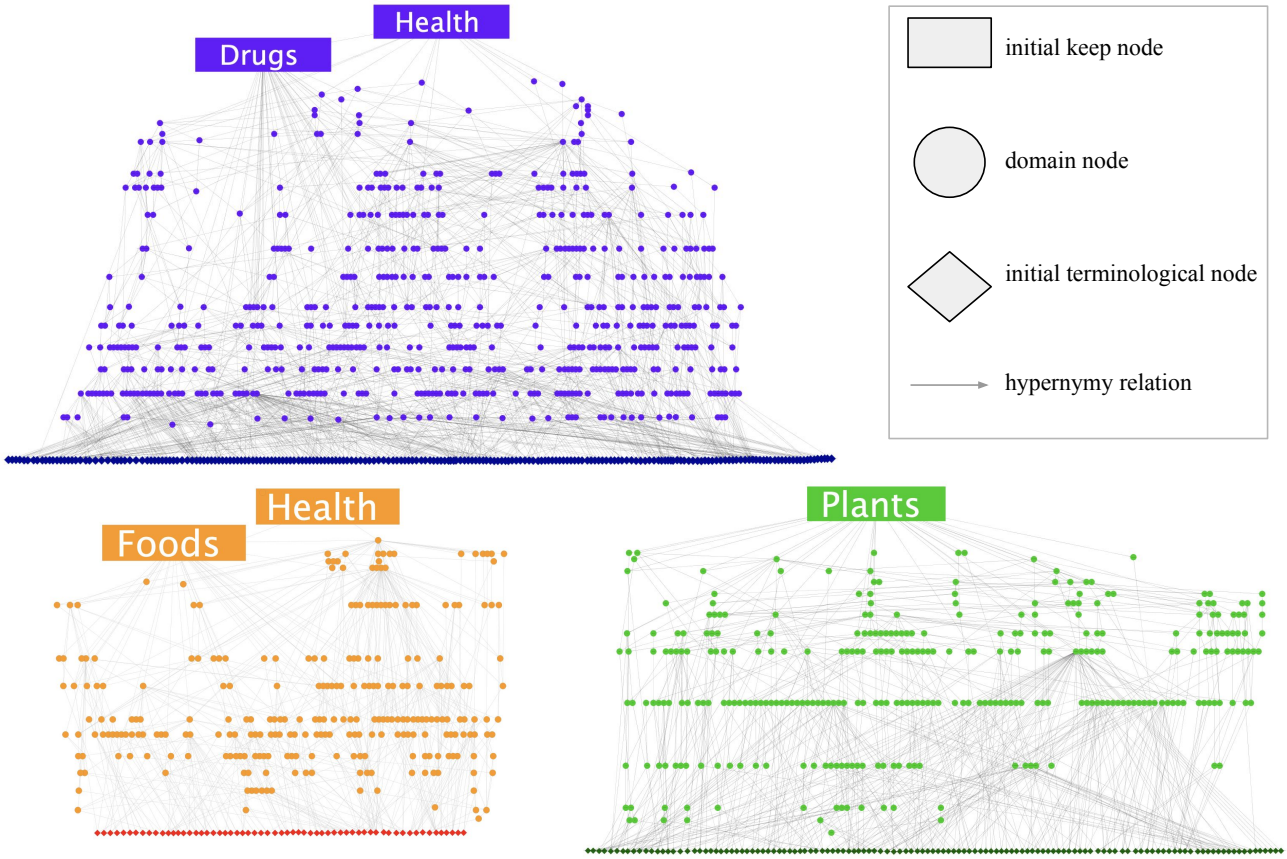

Figure 5: Gold standard DRUGS, FOODS and PLANTS taxonomies.

noisy taxonomy I in a top-down fashion, and cumulatively measures the cluster similarities between the clusters C_G, C_I induced by the connectivity of two adjacent levels of G and of I[6].

The above measures are included in the evaluation frameworks of the two editions of the Taxonomy Evaluation task (TExEval), namely SemEval 2015 - task 17 (Bordea et al., 2015) and SemEval 2016 - task 13 (Bordea et al., 2016).

5. Conclusions and discussion

In this work, we introduced an evaluation framework enabling the comparison against gold/silver standard domain-specific taxonomies extracted from the Wikipedia knowledge graph. The framework includes three manually collected application-specific taxonomies, being DRUGS, FOODS and PLANTS, that are useful for the task of extracting food-drug interactions and herb-drug interactions. Based on our analysis, we conclude that current state-of-the-art algorithms for pruning large knowledge graphs still require a large amount of manual work when applied to broad, real-world application domains such as the ones considered in this work. To address this issue, we proposed an iterative filtering methodology that reduces the amount of manual work for taxonomy verification by a wide margin. Our benchmark constructed using the methodology described in section 2. can be adapted to other domains of interest with minimal effort. In particular, the string-based filtering approach requires domain adaptation but this can be achieved by analysing a reasonable amount of categories. A limitation of our evaluation approach for the gold standard taxonomies is that it does not allow us to calculate recall. To address this, we empirically observed that relevant nodes are typically grouped within the same community and we made use of a border community sampling approach to evaluate nodes that are placed in the immediate vicinity of the main relevant community. This allows us to focus on closely related nodes to ensure completeness as much as possible, with the assumption that nodes that are further away in the graph are less likely to be relevant. Overall, the iterative combination of bottom-up pruning and manual assessment of keep and discard nodes (as described in section 2.) leads to a good coverage of domain relevant nodes. Our further studies will include the development of similar methodologies of extraction and evaluation of multi-domain/application-specific faceted taxonomies (Sacco, 2007), as well as the extraction of taxonomy backbones from large scale noisy hypernymy graphs (Faralli et al., 2019). Faceted taxonomies should reduce the amount of manual work needed for filtering the pruned graph allowing users to keep or discard taxonomy facets as required by the application. This would allow users to discard more easily taxonomy facets related to geographical locations for example, without having to analyse and discard each node one by one.

[6]Given two adjacent levels L_1 and L_2 of a taxonomy, a soft partition of the nodes of L_2 is induced by aggregating nodes sharing the same hypernym

Resource availability

We release the three taxonomies and the annotations of node domain relevance collected for the evaluation (see section 4.). The resource is publicly available at `https://sites.google.com/unitelmasapienza.it/wikipediataxonomies/` under a CC BY 4.0 Licence.

Acknowledgements

This work was supported by the MIAM project and Agence Nationale de la Recherche through the grant ANR-16-CE23-0012 France and by the kANNa project and the European Commission through grant H2020 MSCA-IF-217 number 800578.

Bibliographical References

Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. In *Third international AAAI conference on weblogs and social media*.

Baxter, K. and Preston, C. L. (2010). *Stockley's drug interactions*, volume 495. Pharmaceutical Press London.

Baxter, K., Driver, S., and Williamson, E. (2013). *Stockley's herbal medicines interactions*. Pharmaceutical Press.

Bordea, G., Buitelaar, P., Faralli, S., and Navigli, R. (2015). Semeval-2015 task 17: Taxonomy extraction evaluation (texeval). In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 902–910.

Bordea, G., Lefever, E., and Buitelaar, P. (2016). Semeval-2016 task 13: Taxonomy extraction evaluation (texeval-2). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1081–1091.

Bordea, G., Thiessard, F., Hamon, T., and Mougin, F. (2018). Automatic query selection for acquisition and discovery of food-drug interactions. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 115–120. Springer.

Bordea, G., Randriatsitohaina, T., Mougin, F., Grabar, N., and Hamon, T. (2019). Query selection methods for automated corpora construction with a use case in food-drug interactions. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 115–124.

Camacho-Collados, J., Bovi, C. D., Anke, L. E., Oramas, S., Pasini, T., Santus, E., Shwartz, V., Navigli, R., and Saggion, H. (2018). Semeval-2018 task 9: hypernym discovery. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 712–724.

Faralli, S., Finocchi, I., Ponzetto, S. P., and Velardi, P. (2018a). Crumbtrail: An efficient methodology to reduce multiple inheritance in knowledge graphs. *Knowledge-Based Systems*, 151:180–197.

Faralli, S., Finocchi, I., Ponzetto, S. P., and Velardi, P. (2018b). Efficient pruning of large knowledge graphs. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4055–4063.

Faralli, S., Finocchi, I., Ponzetto, S. P., and Velardi, P. (2019). Webisagraph: A very large hypernymy graph from a web corpus. In *Proceedings of the 6th Italian Conference on Computational Linguistics, Bari, Italy, November 13-15, 2019*.

Fawei, B., Pan, J. Z., Kollingbaum, M., and Wyner, A. Z. (2019). A semi-automated ontology construction for legal question answering. *New Generation Computing*, 37(4):453–478, Dec.

Hulpus, I., Hayes, C., Karnstedt, M., and Greene, D. (2013). Unsupervised graph-based topic labelling using dbpedia. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 465–474.

Jurgens, D. and Pilehvar, M. T. (2016). Semeval-2016 task 14: Semantic taxonomy enrichment. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1092–1102.

Kapanipathi, P., Jain, P., Venkataramani, C., and Sheth, A. (2014). User interests identification on twitter using a hierarchical knowledge base. In *European Semantic Web Conference*, pages 99–113. Springer.

Ojokoh, B. and Adebisi, E. (2019). A review of question answering systems. *Journal of Web Engineering*, 17:717–758, 01.

Sacco, G. (2007). Research results in dynamic taxonomy and faceted search systems. In *8th International Conference on Database and Expert Systems Applications (DEXA 2007)*, pages 201–206, 10.

Velardi, P., Navigli, R., Faralli, S., and Ruiz-Martínez, J. M. (2012). A new method for evaluating automatically learned terminological taxonomies. In *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, pages 1498–1504.