A New Method for Evaluating Automatically Learned Terminological Taxonomies

Paola Velardi†, Roberto Navigli†, Stefano Faralli†, Juana Ruiz Martinez‡

†Dipartimento di Informatica, ‡Facultad de Informàtica
†Sapienza Università di Roma, Italy, ‡Campus de Espinardo, Murcia, Spain
{velardi,navigli,faralli}@di.uniroma1.it, jmruymar@um.es

Abstract

Evaluating a taxonomy learned automatically against an existing gold standard is a very complex problem, because differences stem from the number, label, depth and ordering of the taxonomy nodes. In this paper we propose casting the problem as one of comparing two hierarchical clusters. To this end we defined a variation of the Fowlkes and Mallows measure (Fowlkes and Mallows, 1983). Our method assigns a similarity value $B^i_{(l,r)}$ to the learned (l) and reference (r) taxonomy for each cut i of the corresponding anonymised hierarchies, starting from the topmost nodes down to the leaf concepts. For each cut i, the two hierarchies can be seen as two clusterings C^i_l , C^i_r of the leaf concepts. We assign a prize to early similarity values, i.e. when concepts are clustered in a similar way down to the lowest taxonomy levels (close to the leaf nodes). We apply our method to the evaluation of the taxonomy learning methods put forward by Navigli et al. (2011) and Kozareva and Hovy (2010).

Keywords: taxonomy learning, gold standard evaluation, hierarchical cluster

1. Introduction

Ontology evaluation is a hard task that is difficult even for humans. One reason is that different taxonomies might model the domain of interest equally well. Nonetheless, in the literature a variety of different methods have been proposed for evaluating the quality of a taxonomy. These include (Brank et al., 2006; Maedche et al., 2002): i) manual evaluation performed by domain experts, ii) structural evaluation of the taxonomy, iii) automatic evaluation against a gold standard, iv) application-driven evaluation, in which a taxonomy is assessed on the basis of the improvement its use generates within an application. Other quality indicators have also been analysed in the literature, such as accuracy, completeness and consistency (Volker et al., 2010), and more theoretical features such as essentiality, rigidity and unity (Guarino and Welty, 2002). As regards lexicalised taxonomies, the focus of interest in this paper, the most popular approach (adopted e.g. by Snow et al. (2006), Yang and Callan (2009) and Kozareva and Hovy (2010)) is that of attempting to reconstruct an existing taxonomy (Maedche et al., 2002), like WordNet (Fellbaum, 1998) or the Open Directory Project[1]. This method is applicable when the set of taxonomy concepts is given and the evaluation task is restricted to measuring the ability to reproduce hypernymy links between concept pairs. However, the evalution is far more complex when learning a taxonomy is performed entirely from scratch, as is done by Navigli et al. (2011) and Kozareva and Hovy (2010). In reference taxonomies, the granularity and cotopy (Maedche et al., 2002) of an abstract concept might vary according to the scope of the taxonomy and the expertise of the team who created it. For example, the term *chiaroscuro* is classified under *picture, image, icon* in Wordnet, along with *collage*, but in the Art and Architecture Thesaurus (AA&T)[2] *chiaroscuro* is a *perspective and shading technique*, while *collage* is classified under *image-making processes and techniques*. As long as it is commonsense, non-specialised knowledge that is being considered, it is still feasible for an automated system to replicate an existing classification, because the Web will provide abundant evidence for it. For example, Kozareva and Hovy (2010) are very successful in reproducing the WordNet sub-taxonomy for *animals*, since dozens of definitional patterns are to be found on the Web that classify, e.g., *lion*, as either a *carnivorous feline mammal*, or *carnivorous*, or *feline*. Instead, reconstructing an existing taxonomy in more technical domains is almost impossible, as can be inferred from the foregoing AA&T example.

To tackle this problem we here propose a novel procedure for evaluating a taxonomy against a gold standard, based on reformulating the problem in terms of comparison between hierarchical clusters (cf. Section 2.). To this end the non-leaf concepts of the learned and reference taxonomies are labelled with the transitive closure of their hyponym relations. The procedure is then applied to the task of comparing four automatically acquired taxonomies with the corresponding reference taxonomies (cf. Section 3.), namely: the virology sub-hierarchy of the MeSH[3] medical taxonomy, and three sub-hierarchies of WordNet[4] (animal, plants and vehicles). In Section 4. we compare the proposed methodology with other approaches in the literature. Finally, Section 5. is dedicated to concluding remarks.

2. Evaluation Method

In this section we propose a novel, general measure for the evaluation of a learned taxonomy against a gold standard. We borrow Brank et al.'s (2006) idea of exploiting the analogy with unsupervised clustering but, rather than representing the two taxonomies as flat clusterings (see Section

[1] http://www.dmoz.org/

[2] http://www.getty.edu/vow/AATHierarchy

[3] http://www.nlm.nih.gov/mesh/

[4] http://wordnet.princeton.edu/

4.), we propose a measure that takes into account the hierarchical structure of the two taxonomies being analyzed. From this perspective, a taxonomy can be transformed into a hierarchical clustering by replacing each label of a non-leaf vertex, e.g., *perspective and shading techniques*, with the transitive closure of its hyponyms, e.g., *cangiatismo, chiaroscuro, foreshortening, hatching*.

Techniques for comparing clustering results have been surveyed by Wagner and Wagner (2007). However, to the best of our knowledge, the only method for comparing hierarchical clusters is that proposed by Fowlkes and Mallows (1983). Suppose that we have two hierarchical clusterings H_1 and H_2, with an identical set of *n* objects. Let k be the maximum depth of both H_1 and H_2, and H_j^i a cut of the hierarchy, where $i \in \{0, \ldots, k\}$ is the cut level and $j \in \{1, 2\}$ selects the clustering of interest. Then, for each cut i, the two hierarchies can be seen as two flat clusterings C_1^i and C_2^i of the n concepts. When $i = 0$ the cut is a single cluster incorporating all the objects, and when $i = k$ we obtain n singleton clusters. Now let:

- n_{11} be the number of object pairs that are in the same cluster in both C_1^i and C_2^i;
- n_{00} be the number of object pairs that are in different clusters in both C_1^i and C_2^i;
- n_{10} be the number of object pairs that are in the same cluster in C_1^i but not in C_2^i;
- n_{01} be the number of object pairs that are in the same cluster in C_2^i but not in C_1^i;

The generalized Fowlkes and Mallows (F&M) measure of cluster similarity for the cut i ($i \in \{0, \ldots, k-1\}$), as reformulated by Wagner and Wagner (2007), is defined as:

$$B_{1,2}^i = \frac{n_{11}^i}{\sqrt{(n_{11}^i + n_{10}^i) \cdot (n_{11}^i + n_{01}^i)}}. \quad (1)$$

Note that the formula can be interpreted as the geometric mean of precision and recall of an automated method of clustering the same concept pairs as in a gold-standard clustering. However, this formula has a few inconvenient properties: first, the value of $B_{1,2}^i$ gets close to its maximum 1.0 as we approach the root of the hierarchy ($i = 0$); second, the two hierarchies need to have the same maximum depth k; third, the hierarchies need to have the same number of initial objects and a crisp classification.

In order to apply the F&M measure to the task of comparing a learned and a gold-standard taxonomy we need to mitigate these problems. Formula 1 allows the measure to cope with the third problem without modifications. In fact, if the sets of objects in H_1 and H_2 are different, the integers n_{10} and n_{01} can be considered as also including objects that belong to one hierarchy and not to the other. In this case, the value of $B_{1,2}^0$ will provide a measure of the objects in common between the learned taxonomy and the gold-standard one. To account for multiple (rather than crisp) classifications, again, there is no need to change the formula, which is still relevant if an object is allowed to belong to more than one cluster. As before, mismatches between H_1 and H_2 would result in higher values of n_{10} and n_{01} and lower $B_{1,2}^i$.

A more serious problem with formula 1 is that the lower the value of *i*, the higher the value of the formula, whereas, ideally, we would like to reward similar clusterings when the clustering task is more difficult and fine-grained, that is, for cuts that are close to the leaf nodes. To assign a reward to "early" similarity values, we weight the values of $B_{1,2}^i$ with a coefficient $\frac{i+1}{k}$. We can then compute a cumulative measure of similarity with the following formula:

$$B_{1,2} = \frac{\sum_{i=0}^{k-1} \frac{i+1}{k} B_{1,2}^i}{\sum_{i=0}^{k-1} \frac{i+1}{k}} = \frac{\sum_{i=0}^{k-1} \frac{i+1}{k} B_{1,2}^i}{\frac{k+1}{2}}. \quad (2)$$

Finally, to solve the problem of different depths of the two hierarchies we define a policy that penalizes a learned taxonomy that is less structured than the gold-standard one, and rewards – or at least does not penalize – the opposite case.

As an example, consider Figure 1, showing two taxonomies H_1 and H_2, with non-identical sets of objects $\{a, b, c, d, e, f\}$ and $\{a, b, c, d, e, g\}$. In the figure, each edge is labeled by its distance from the root node (the value i in the F&M formula). Notice that H_1 and H_2 have multiple classifications (i.e., multiple hypernyms in our case) for the object *e*, thus modelling the common problem of lexical ambiguity and polysemy. Let us suppose that H_1 is the learned taxonomy, and H_2 the gold-standard one. We start comparing the clusterings at cut 0 and stop at cut $k_r - 1$, where k_r is the depth of the gold-standard taxonomy. This means that if the learned taxonomy is less structured we replicate the cut k_l for $k_r - k_l$ times, while if it is more structured, we stop at cut k_l (k_l is the maximum depth of the learned taxonomy). In contrast to previous evaluation models, our aim is to reward (instead of penalize) more structured taxonomies provided they still match the gold standard one.

Table 1 shows the cuts from 0 to 3 of H_1 and H_2 and the values of $B_{1,2}^i$. For $i = 2$ the B value is 0 if H_2 is the learned taxonomy, and is not defined if H_2 is the gold standard. Therefore, when computing the cumulative formula 2, we obtain the desired effect of penalising less structured learned taxonomies. Note that when the two hierarchies have different depth, the value $k - 1$ in formula 2 is replaced by $k_r - 1$.

3. Experiments

In this section we apply our modified F&M evaluation model to the comparison of a taxonomy learned from scratch using the ontology learning methodology described by Navigli et al. (2011), against the following gold standards:

- three WordNet sub-hierarchies, namely *animals, vehicles* and *plants*. We selected these three WordNet sub-hierarchies to enable a comparison to be made with Kozareva and Hovy's (2010) taxonomy learning method, which was also tested on these domains;
- The virology sub-hierarchy of MeSH.

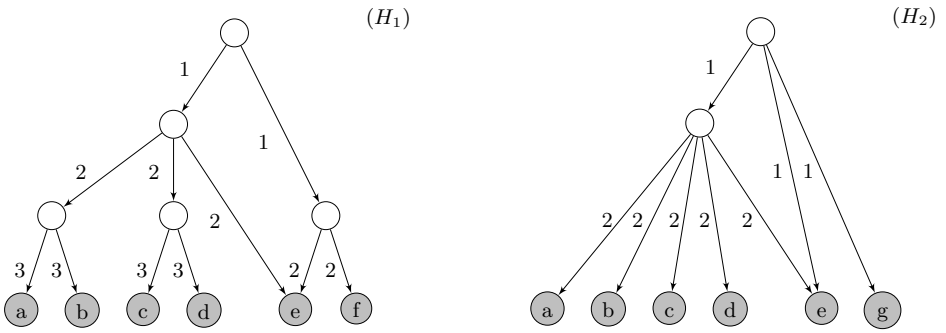

Figure 1: Two hierarchical clusters of *n* non-identical objects.

i	C_1	C_2	n_{11}	n_{10}	n_{01}	H_1	H_2 $B^i_{1,2}$
0	{a,b,c,d,e,f}	{a,b,c,d,e,g}	10	5	5	0.67	0.67
1	{a,b,c,d,e},{e,f}	{a,b,c,d,e},{e},{g}	10	1	0	0.95	0.95
2	{a,b},{c,d},{e},{f}	{a},{b},{c},{d},{e},{g}	0	2	0	undefined	0
3	{a},{b},{c},{d},{e},{f}	{a},{b},{c},{d},{e},{g}	0	0	0	undefined	undefined

Table 1: Application of the evaluation method to the hierarchies of Figure 1. The values of $B^i_{1,2}$ are shown both when H_1 and H_2 are the learned taxonomy (penultimate and last column, respectively).

We first start with a summary description of the ontology learning algorithms proposed in (Navigli et al., 2011) and (Kozareva and Hovy, 2010). Then, we compare the learned taxonomies againts the reference ones, using formulas (1) and (2). This analysis proves particularly helpful in pinpointing some significant and recurrent phenomena. Finally, we perform an *in-vitro* assessement of the method – as was done in (Zavitsanos et al., 2011) and (Brank et al., 2006) – introducing artificial modifications (concept swapping, adding/deleting intermediate nodes, etc.) into reference taxonomies, in order to evaluate their effect on the proposed similarity measure.

3.1. Summary of analyzed taxonomy learning methods

3.1.1. OntoLearn Reloaded

In (Navigli et al., 2011) we presented a system, named OntoLearn Reloaded, whose objective was to produce a domain taxonomy starting only from a domain corpus and the Web. To this end we devised a methodology based on two "core" algorithms for definition extraction and graph-based taxonomy induction, empowered by some pre- and post-processing steps. Our graph-based taxonomy induction approach consists of five steps, as shown in Figure 2. We start from an initially-empty directed graph $G_{noisy} = (V_{noisy}, E_{noisy})$, where $V_{noisy} := \emptyset$ and $E_{noisy} := \emptyset$. From a corpus of domain documents we acquire a domain terminology T, using a terminology extraction tool. For all the terms in T, we mine the corpus and the Web in search of definitional expressions, from which we obtain a set of hypernyms H using a lattice based definition classifier and hypernym extractor (Navigli and Velardi, 2010).
Non-domain definitions are eliminated on the basis of the number of domain-related words in the definition. The retrieved hypernyms are then used for a new Web search, and the process iterates until a termination condition is satisfied. The result is a highly dense hypernym graph with several cycles and possibly disconnected components. Over this directed graph, we apply an algorithm for finding an optimal branching. Finally, since in many practical applications DAGs (directed acyclic graphs) represent a more appropriate abstraction than tree-like taxonomies, we apply an edge recovery strategy to re-attach some of the hypernym edges deleted during the optimal branching step.
In our previously published experiments we applied OntoLearn Reloaded to the task of acquiring a brand new taxonomy for the domain of Artificial Intelligence, for which we performed a manual assessment since no gold standard taxonomies were available for this domain. In this paper, instead, we apply OntoLearn Reloaded to the task of inducing four taxonomies for which a gold standard reference is available. We evaluate three outputs of the OntoLearn algorithm: $TREE$, $DAG[1-3]$ and $DAG[0-99]$, corresponding respectively to the output obtained after optimal branching, and to the DAGs obtained with two different edge recovery policies, of which the first is more conservative (Velardi et al., 2012).

3.1.2. Doubly Anchored Patterns

Kozareva and Hovy (2010)[K&H] create a hypernym graph in three steps. Given few root concepts (e.g., ANIMAL) and basic level concepts or instances (e.g., *lion*), they:

1) harvest new basic and intermediate concepts from the Web in an iterative fashion, using doubly-anchored patterns (DAP) like '$\langle root \rangle$ *such as* $\langle seed \rangle$ *and* $*$' and inverse DAP (i.e., DAP^{-1}) like '$*$ *such as* $\langle term1 \rangle$ *and* $\langle term2 \rangle$';

2) rank the extracted DAP and DAP^{-1} nodes by out-degree and in-degree, respectively, to prune out less promising terms;

3) induce the final taxonomic structure by positioning the intermediate nodes betwen basic level and root terms

using a variety of surface patterns along the line of Hearst (1992).

4) finally, they eliminate cycles, as well as nodes with no predecessor or no successor, and they select the longest path in case of multiple paths between node pairs.

As remarked in the introduction, K&H do not actually apply their algorithm to the task of creating a new taxonomy, but instead try to reproduce three WordNet taxonomies, under the assumption that the taxonomy nodes are known. In other terms, when given a reference taxonomy, during the graph growing steps 1 and 2, they reject nodes that do not belong to the reference.

3.2. Results

Figure 3 shows, for each domain ((a) ANIMALS, (b) PLANTS, (c) VEHICLES, and (d) VIRUSES), and for each "anonymized" hierarchy, a plot of $B^i_{1,2}$ values multiplied by the penalty factor. The generally decreasing values of $B^i_{1,2}$ in Figure 3 show that, as expected, mimicking the clustering criteria of a taxonomy created by a team of experts proves very difficult at the lowest levels, while performance grows in the subsequent generalization steps. As far as the comparison with K&H is concerned, we note that, though K&H obtain in general better performance[5], OntoLearn Reloaded has higher coverage over the domain, as is shown by the highest values for $i = 0$, and, especially with $DAG[0-99]$, has a higher depth of the derived hierarchy. Another recurrent phenomenon is that K&H curves gracefully degrade from the root to the leaf nodes, possibly with a peak in the intermediate levels, while OntoLearn Reloaded has a hollow in the mid-high region (see the region 4-6 for ANIMALS and 1-2 for the other three hierarchies) and often a relative peak in the lowest levels. In what follows we explain these recurrent phenomena.

As we move from leaf nodes to the upper ontology, the extracted terms become progressively more general and consequently more ambiguous. OntoLearn uses a context-based disambiguation strategy which is rather successful at the lowest levels, but is more error-prone when moving towards the upper levels. But why are these errors frequent at the intermediate levels and not at the highest levels? An example in the ANIMAL domain is represented by the induced hypernymy sequence $fawn \leftarrow color \leftarrow race \leftarrow breed \leftarrow domestic\ animal$, where the wrong hypernym *color* was originated by the definition *"FAWN is a light yellowish brown COLOR that is usually used in reference to a dog's coat color."*. Here, the word "dog" caused the sentence to be considered in-domain. In many cases, wrong hypernyms do not accumulate sufficient weight and create "dead-end" hypernymy chains, which are pruned during the optimal branching step (see Section 3.1.), but unfortunately a domain appropriate definition is found for *color*: *"a COLOR is a RACE with skin pigmentation different from the white race"*. On the other hand, this new sentence produces an attachment that, in a sense, rectifies the error, because *race* is a "good" domain concept that eventually ends up in subsequent iterations to the upper node *domestic animal*. Many examples like this can be found in all domains. Clearly, K&H do not experience the same phenomenon, since they assume that the set of ontology concepts is known and reject any non gold standard node.

At the lowest taxonomy levels, errors are caused by two contrary phenomena: overgeneralization and overspecialization. For example, *macaque* has *monkey* as a direct hypernym in WordNet, while we find *short-tailed monkey* as a direct hypernym of *macaque*. An opposite case is *ganoid* which is a *taleostan* in WordNet and simply a *fish* in our taxonomy. The first case does not reward the learned taxonomy (though, unlike for the *overlapping factor* (Maedche et al., 2002), it does not cause a penalty), while the second is quite penalizing.

Finally, in Table 2 we show the cumulative $B_{1,2}$ values for the four domains, according to formula 2. Here, except for the VEHICLES domain, the unconstrained $DAG[0, 99]$ performs best.

Experiment	VIRUSES	ANIMALS	PLANTS	VEHICLES
TREE	0.123	0.102	0.174	0.092
DAG [1,3]	0.139	0.091	0.217	0.094
DAG [0,99]	**0.165**	**0.162**	**0.251**	0.132
K&H	n.a.	0.11	0.229	**0.518**

Table 2: Values of $B_{1,2}$ for the domains of VIRUSES, ANIMALS, PLANTS, VEHICLES.

3.3. Artificial Evaluation

In (Zavitsanos et al., 2011) and (Brank et al., 2006) the evaluation methodology is assessed in a controlled experimental setting, by using a set of damage operators on a reference gold standard taxonomy. In this Section we apply the same assessment technique on the four taxonomies of Figure 3 and on the GENIA ontology[6]. We apply the same damage operators introduced in (Zavitsanos et al., 2011), namely: swap concept, remove concept, add concept, add relation. We did not include the operator *change concept representation* because it only applies to their methodology, summarized in Section 4..

For each damage type, we apply a growing perturbation degree, from 10% to 100%, to monitor the system behaviour under realistic conditions, since rarely in the literature does the overlapping factor between a learned and a reference taxonomy exceed 50%. To compare with (Zavitsanos et al., 2011), for the GENIA ontology we also applied a different strategy with only 10 perturbations. For each perturbation, we ran 50 different tests, thus performing 500 tests per damage type as was done in (Zavitsanos et al., 2011).

When comparing the two measures, see Figures 5 (a) and (b), it is interesting to observe that only the "add relation" perturbation produces more or less the same effect. Viceversa, the comparison shows that the F&M evaluation model is less penalizing for "add concept" and "swap concept",

[5] Again we remark that K&H, when reproducing a reference taxonomy, reject nodes that do not belong to the reference, while OntoLearn Reloaded is only provided with a set of "leaf" nodes. In this experiment, we used the same set of seeds as in (Kozareva and Hovy, 2010)

[6] http://www.tsujii.is.s.u-tokyo.ac.jp/GENIA

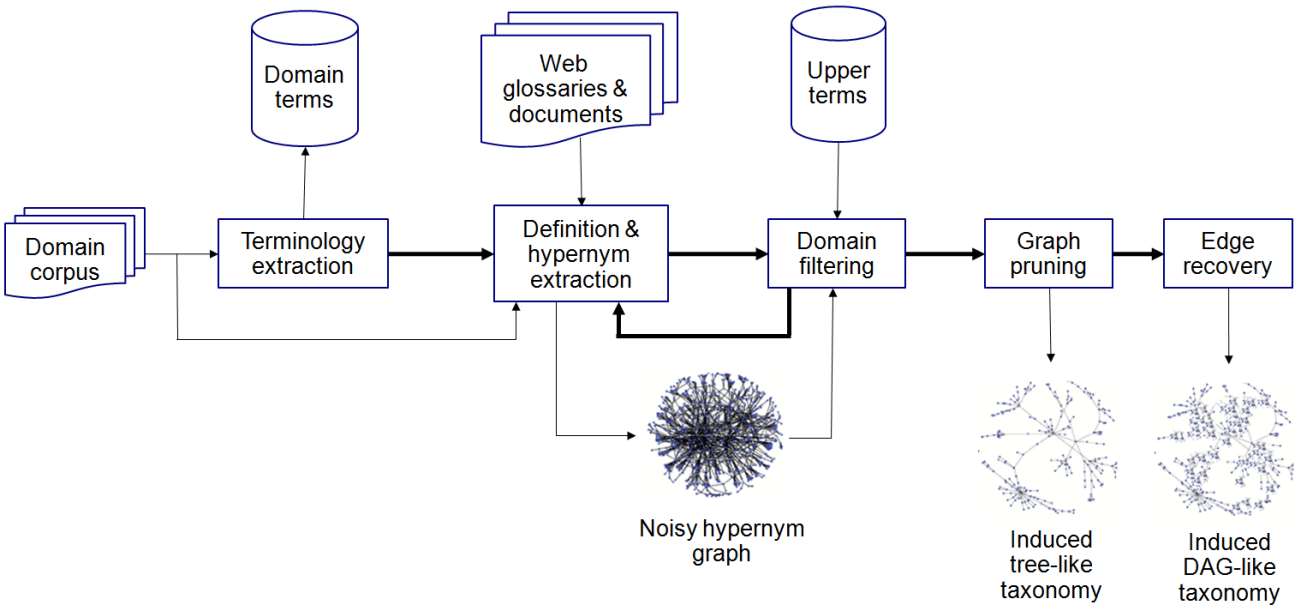

Figure 2: The taxonomy learning workflow.

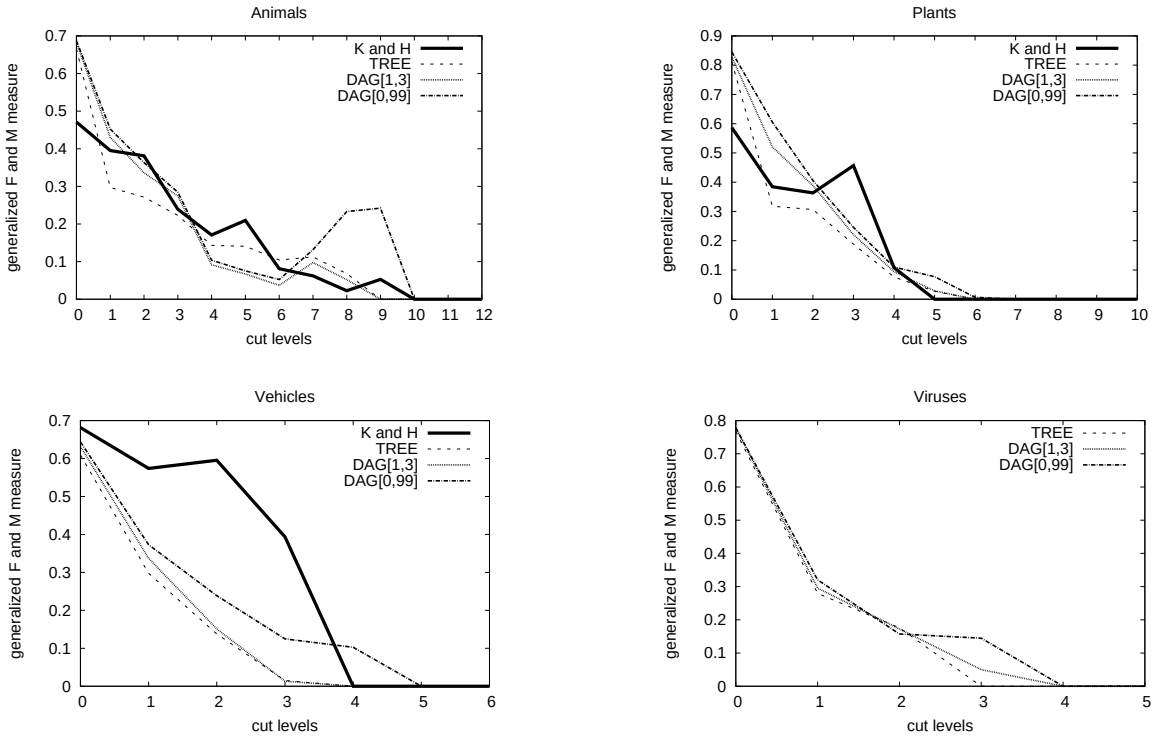

Figure 3: Gold standard evaluation of our three versions of OntoLearn Reloaded against WordNet (Animals, Plants and Vehicles) and MeSH (Viruses). A comparison with K&H is also shown for the first three domains.

and more so for "remove concept". This is a desired effect, because discovering new concepts does not necessarily mean producing an error (remember, e.g., the *short-tailed monkey* example), while removing concepts certainly implies a loss of information. Finally the global damage of a swap is reduced in our F&M model, since the hierarchical clustering methodology eventually cancels the error (soon or later the swapped concepts rejoin).

4. Related Work

Gold Standard evaluation against an automatically learned ontology has been analyzed in a systematic way in (Zavitsanos et al., 2011) and (Brank et al., 2006). Both methods attempt to escape the naming problem that we outlined in the introduction, adopting two different strategies. Zavistanos and his colleagues (Zavitsanos et al., 2011) propose transforming the ontology concepts and their properties into distributions over the terms space of the source data from which the ontology has been learned. These dis-

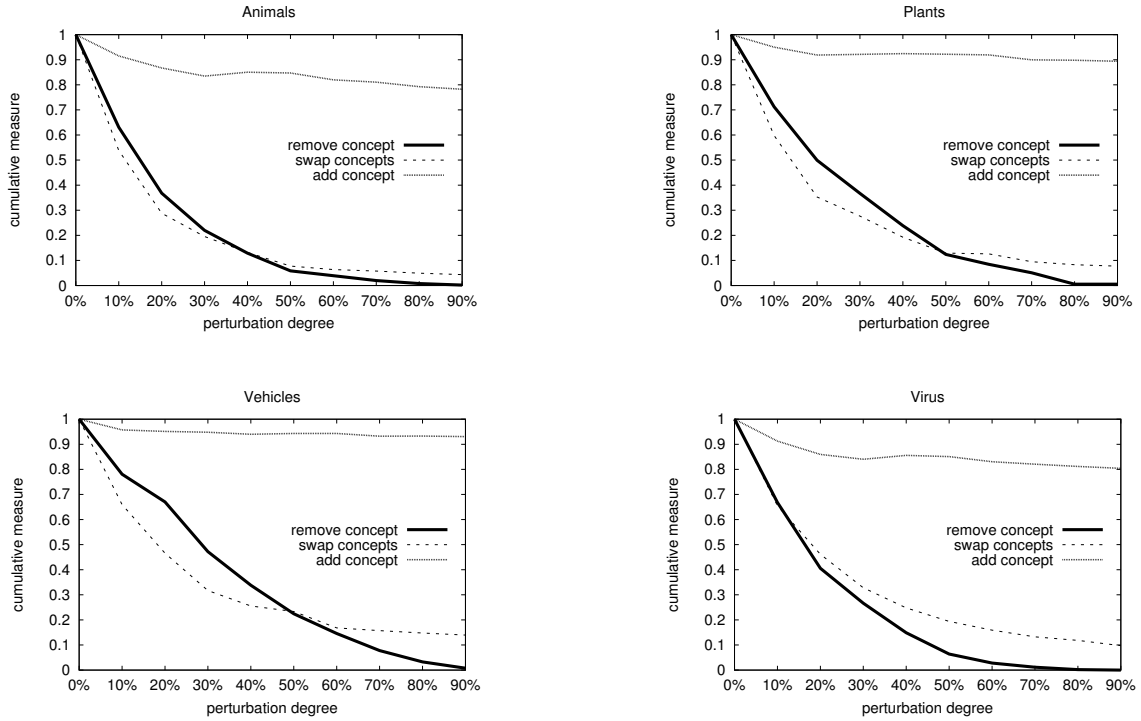

Figure 4: Artificial perturbations on the Animals Plants Vehicles and Viruses gold standards.

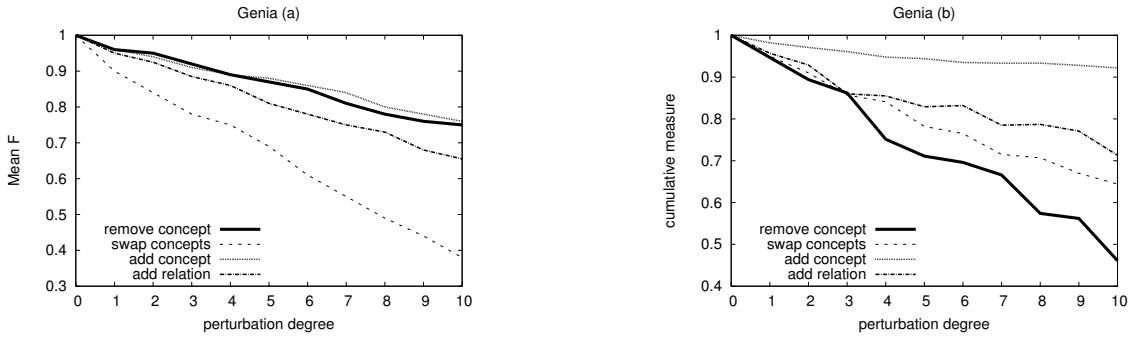

Figure 5: Comparison between the method in (Zavitsanos et al., 2011) (a) and the modified F&M measure (b).

tributions are used to compute pairwise concept similarity between reference and learned ontologies. The approach seems interesting, however, it is evaluated only within an *in-vitro* experiment, here described in Section 3.3..

Brank et al. (2006) exploit the analogy between ontology learning and unsupervised clustering, as we do, and propose the Rand Index (Rand, 1971) to compute the similarity between a learned ontology l and a gold standard r. The Rand Index measures the similarity between two clusterings C_l and C_r by the formula: $R(Cl, Cr) = \frac{2(n_{11}+n_{00})}{n(n-1)}$ where n_{11}, n_{00} and n have the same meaning as for the F&M measure. R ranges from 0 (no pair classified in the same way under both clusterings) to 1 (identical clusterings). In (Brank et al., 2006), a clustering is obtained from an ontology by associating every ontology instance to its concept. The set of clusters is hence represented by the set of leaf concepts in the hierarchy, i.e. , according to our notation, the clustering C_i^{k-1}. To account for the hierarchical structure, they define the OntoRand formula. This measure, rather than returning 1 or 0 depending whether or not two given instances i and j belong to the same cluster in the compared ontologies, returns a real number in $[0, 1]$ depending upon the distance between i and j in terms of common ancestors. In other terms, if i and j do not belong to the same concept but have a very close common ancestor, the OntoRand measure still returns a value close to 1. Morey and Agresti (1981) demonstrated a high dependency of the Rand Index upon the number of clusters, while Fowlkes and Mallows (1983) show that the Rand Index has the undesirable property of converging to 1 as the number of clusters increases, even in the unrealistic case of independent clusterings. These undesired outcomes have been experienced also in (Brank et al., 2006), who note in their experimental section that: *the similarity of an ontology to the original one is still 0.74 even if only the top three levels of the ontology have been kept*. Another problem with the OntoRand formula, as also remarked in (Zavitsanos et al., 2011), is the requirement of comparing ontologies with the

same set of instances.

To summarize, our modified F&M measure has several advantages over previous approaches, since:

i) it allows comparison at different levels of depth of the hierarchy, penalizing errors at the highest cuts of the hierarchy;

ii) it does not require that the two hierarchies have the same depth, nor that they have the same number of leaf nodes;

iii) the measure can be extended to lattices, e.g. it is not required that each object belongs precisely to one cluster;

iv) it penalizes the introduction of new concepts (which cannot be a-priori considered as errors) to a lesser degree and assigns a graded penalty to concept swaps, depending on their hierarchical distance in the learned taxonomy.

5. Conclusions

In this paper we proposed a methodology for evaluating a taxonomy learned entirely from scratch against an existing gold standard. The method has some advantages over existing approaches, such as (Zavitsanos et al., 2011) and (Brank et al., 2006). Our methodology was assessed both through the evaluation of four automatically learned taxonomies and through artificial perturbation experiments.

Acknowledgments

Roberto Navigli and Stefano Faralli gratefully acknowledge the support of the ERC Starting Grant MultiJEDI No. 259234.

6. References

J. Brank, D. Mladenic, and M. Grobelnik. 2006. Gold standard based ontology evaluation using instance assignment. In *Proc. Fourth Workshop Evaluating Ontologies for Web EON 2006*.

C. Fellbaum, editor. 1998. *WordNet: An Electronic Database*. MIT Press, Cambridge, MA.

E.B. Fowlkes and C.L. Mallows. 1983. A method for comparing two hierarchical clusterings. In *Journal of the American Statistical Association*, volume 78, pages 553–569.

N. Guarino and C. Welty. 2002. Evaluating ontological decisions with ontoclean. In *Communications of the ACM*, volume 45, pages 61–65.

M.A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proc. of COLING*, pages 539–545.

Z. Kozareva and E. Hovy. 2010. A semi-supervised method to learn and construct taxonomies using the web. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 1110–1118, Cambridge, MA, October.

A. Maedche, V. Pekar, and S. Staab. 2002. Ontology learning part one: On discovering taxonomic relations from the web. In *Proceedings of Web Intelligence*, pages 301–322.

L. C. Morey and A. Agresti. 1981. An adjustment to the rand statistic for chance agreement. In *The Classification Society Bulletin*, volume 5, pages 9–10.

R. Navigli and P. Velardi. 2010. Learning Word-Class Lattices for definition and hypernym extraction. In *Proceedings of ACL*, pages 1318–1327, Uppsala, Sweden.

R. Navigli, P. Velardi, and S. Faralli. 2011. A graph-based algorithm for inducing lexical taxonomies from scratch. In *Proceedings of the* 22^{nd} *International Joint Conference on Artificial Intelligence*, pages 1872–1877, Barcelona, Spain.

W. M. Rand. 1971. Objective criteria for the evaluation of clustering methods. In *Journal of the American Statistical Association*, pages 66(336):846–850.

R. Snow, D. Jurafsky, and A. Ng. 2006. Semantic taxonomy induction from heterogeneous evidence. In *Proc. of COLING-ACL*, pages 801–808.

P. Velardi, S. Faralli, and R. Navigli. 2012. Ontolearn Reloaded: A graph-based algorithm for taxonomy induction. *Submitted to Computational Linguistics*.

J. Volker, J. Vrandecic, J. Sure, and A. Hotho. 2010. Aeon - an approach to the automatic evaluation of ontologies. In *Journal of Applied Ontology*, volume 3, pages 41–62.

S. Wagner and D. Wagner. 2007. Comparing clusterings - an overview. *Analysis*, 4769:1–19.

H. Yang and J. Callan. 2009. A metric-based framework for automatic taxonomy induction. In *Proceedings of ACL*, pages 271–279, Stroudsburg, USA.

E. Zavitsanos, G. Paliouras, and G. A. Vouros. 2011. Gold standard evaluation of ontology learning methods through ontology transformation and alignment. In *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, volume 23, pages 1635–1648.