SemEval-2015 Task 17: Taxonomy Extraction Evaluation (TExEval)

Georgeta Bordea, Paul Buitelaar
Insight
Centre for Data Analytics
National University of Ireland, Galway
name.surname@insight-centre.org

Stefano Faralli, Roberto Navigli
Dipartimento di Informatica
Sapienza University of Rome
Italy
surname@di.uniroma1.it

Abstract

This paper describes the first shared task on Taxonomy Extraction Evaluation organised as part of SemEval-2015. Participants were asked to find hypernym-hyponym relations between given terms. For each of the four selected target domains the participants were provided with two lists of domain-specific terms: a WordNet collection of terms and a well-known terminology extracted from an online publicly available taxonomy. A total of 45 taxonomies submitted by 6 participating teams were evaluated using standard structural measures, the structural similarity with a gold standard taxonomy, and through manual quality assessment of sampled novel relations.

1 Introduction

SemEval-2015 Task 17 is concerned with the automatic extraction of hierarchical relations from text and subsequent taxonomy construction. A taxonomy is a hierarchy of concepts that expresses parent-child or broader-narrower relationships. Because of their many applications in search, retrieval, website navigation, and records management, taxonomies are valuable resources for libraries, publishing companies, online databases, and e-commerce companies. Taxonomies are most often manually created resources that are expensive to construct and maintain, and therefore there is a need for automatic methods for taxonomy enrichment and construction. Recently, the task of taxonomy learning from text, also called taxonomy induction, has received an increased interest in the natural language processing community, as taxonomical information is a valuable input to many semantically intensive tasks including inference, question answering (Harabagiu et al., 2003) and textual entailment (Geffet and Dagan, 2005).

Taxonomy learning can be divided into three main subtasks: term extraction, relation discovery, and taxonomy construction. *Term extraction* is a relatively well-known task, hence we decided to abstract from this stage and provide a common ground for the next steps by making available the list of terms beforehand. Most approaches for *relation discovery* from text rely on lexico-syntactic patterns (Hearst, 1992; Kozareva et al., 2008), co-occurrence information (Sanderson and Croft, 1999), substring inclusion (Nevill-Manning et al., 1999), or exploit semantic relations provided in textual definitions (Navigli and Velardi, 2010). Any asymmetrical relation that indicates subordination between two terms can be considered, but here the focus is mainly on hyponym-hypernym relations. Depending on the approach selected, the task may or may not require large amounts of text to extract relations between terms, therefore no corpus is provided as part of the shared dataset.

This stage usually produces a large number of noisy, inconsistent relations, that assign multiple parents to a node and that contain cycles, i.e., sequences of vertices that start and end at the same vertex. Hence, the third stage of taxonomy learning, *taxonomy construction*, focuses on the overall structure of the resulting graph and aims to organise terms into a hierarchical structure, more specifically a directed acyclic graph (Kozareva and Hovy,

Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 902–910,
Denver, Colorado, June 4-5, 2015.

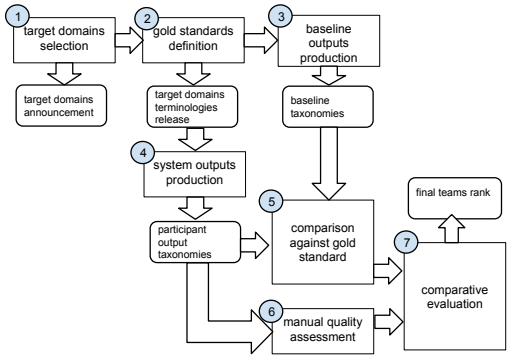

Figure 1: The task workflow.

2010; Navigli et al., 2011; Wang et al., 2013). To address the inherent complexity of evaluating taxonomy quality, several methods have been considered in the past including manual evaluation by domain experts, structural evaluation, and automatic evaluation against a gold standard (Velardi et al., 2012). In this task, all these existing evaluation approaches are considered, using a voting scheme to aggregate the results for the final ranking of the systems. We introduce four new domains that have not previously been considered for this task, covering general knowledge domains such as food and equipment and technical domains such as chemicals and science. For each domain, we provide a gold standard taxonomy gathered exclusively from WordNet (Fellbaum, 2005), as well as a gold standard taxonomy that combines terms and relations gathered from other domain-specific sources.

2 Task workflow

In this section we present the task workflow, the considered dataset, and the evaluation method used in this task.

Competition setup: In order to provide a common ground to all the competing teams, we applied the task workflow described in Figure 1, as follows: 1) select and announce a set of target domains (see Section 2.1 for more details); 2) define and collect gold standard taxonomies that will be used for evaluation and extract and release the set of terms that they cover; 3) select and produce baseline taxonomies using naive baselines to be compared against the team outputs in the competition.

Competition and evaluation flow: As described in Figure 1, the next steps of the workflow concern the participation of the competing teams and the evaluation of the resulting outputs as follows: 4) in this stage participants produce and submit the output taxonomies. For each domain, test data consists of a list of domain terms that participants have to structure into a taxonomy, with the possibility of adding further intermediate terms. Each system will return a list of pairs (term, hypernym). In this way, taxonomy learning is limited to finding relations between pairs of terms and organising them into a hierarchical structure. Participants are encouraged to consider polyhierarchies when organising terms. In this setting, nodes can have more than one parent and the final structure of the taxonomy is not necessarily a tree; 5) compare system outputs (4) and baseline taxonomies (3) with taxonomies produced as gold standards (2); 6) manually annotate a sample of system outputs to estimate the quality of hypernym-hyponym relationships that are not in the gold standards; 7) create a combined rank of the teams based on the individual rank that each team reached on different aspects of the evaluation.

Table 1: Structural measures of Combined and WordNet gold standard taxonomies.

		Combined taxonomies		**WordNet taxonomies**	
Domain	**Root concept**	**|V|**	**|E|**	**|V|**	**|E|**
Chemicals	chemical	17584	24817	1351	1387
Equipment	equipment	612	615	475	485
Food	food	1156	1587	1486	1533
Science	science	452	465	429	441

2.1 Data

We selected four target domains with a rich, deep, hierarchical structure (i.e. Chemicals, Equipment, Food and Science) with four root concepts (i.e. chemical, equipment, food and science, respectively). Then, for each domain we produced two kinds of gold standard taxonomies.

WordNet taxonomy Concepts and relationships in the WordNet hypernym-hyponym hierarchy rooted on the corresponding root concept.

Combined taxonomy Domain-specific terms and relations from well-known, publicly available, tax-

onomies other than WordNet: CheBI[1] for Chemicals, "The Google product taxonomy"[2] for Foods, the "Material Handling Equipment"[3] taxonomy for Equipment, and the "Taxonomy of Fields and their Subfields"[4] for Science. Hypernym-hyponym relationships were also gathered from a general purpose resource, the Wikipedia Bitaxonomy (WiBi) (Flati et al., 2014), using a semi-automatic approach. For each domain we first manually identified domain sub-hierarchies from WiBi (W); Second we automatically searched for the terms of W in common with the corresponding gold standard G. For each common term t we added in G the taxonomy rooted on t from W.

Table 1 shows the resulting number of vertices $|V|$, i.e., the number of terms given to the participants, and the number of edges $|E|$ of the produced gold standard taxonomies for the four target domains. Finally, test data consists of eight lists of domain concepts, for which participants were asked to output a set of hypernym-hyponym relationships.

2.2 Evaluation method

Let $S = (V_S, E_S)$ be an output taxonomy produced by a system for a given domain, where V_S includes the set of domain concepts initially provided by the task organisers and E_S is the set of taxonomy edges extracted by the system. To broadly analyze the quality of the produced set of hypernymy relationships E_S, these results are benchmarked against two naive baselines, described in Section 2.2.1, using the following evaluation approaches: i) analyse the graph structure and check if the produced taxonomy is a Directed Acyclic Graph (DAG); ii) compare the edges E_S, against the set of relations from each type of gold standard; iii) manually validate a sample of novel relationships produced by the system that are not contained in the gold standard.

The final ranking of the systems takes into consideration these three types of evaluation by aggregating the achieved ranks using a voting scheme. First, the output taxonomies are ranked on the basis of the average performance obtained for each evaluated aspect and for each domain. The resulting ranks are simply summed up, favouring systems at the top of the ranked list and penalising systems at the lower end.

2.2.1 Baselines

The main purpose of introducing the baselines described in this section is to check the performance of a system that relies mainly on the fact that the root of the domain is known and implements simple string-based approaches. In this task, the following two naive approaches for taxonomy construction are implemented and used for benchmarking systems:

Baseline 1 Simply connect all the nodes to the root concept: $B_1 = (V_{B_1}, E_{B_1})$ where $E_{B_1} = \{(root, a), a \in V_{B_1} \setminus \{root\}\}$;

Baseline 2 A basic string inclusion approach that covers relations between compound terms such as *(science, network science)*: $B_2 = (V_{B_2}, E_{B_2})$ where $E_{B_2} = \{(a, b)$, b starts with a or ends with a and $|b| > |a|\}$, and where a is a term and b is a compound term that includes a as a substring.

Both approaches require only the root of the taxonomy and the list of terms and do not require any external corpora or other structured information.

2.2.2 Structural analysis

The main goal of the structural evaluation of a taxonomy is to quantify the size of the taxonomy under investigation in terms of nodes and edges. A second objective is to evaluate whether the overall structure connects all the nodes in the graph with the root and whether it is consistent with the semantics of the ISA relation. Hierarchical relations are generally inconsistent with the presence of cycles. Also, we highlight the number of nodes located on higher levels of a taxonomy, called intermediate nodes. These nodes are considered more important than leaves, to favour taxonomies with a deep, rich structure.

Based on these considerations, structural evaluation is performed by computing the cardinality of $|V_S|$ and $|E_S|$. A topological sorting-based algorithm (Kahn, 1962) is used to establish if the taxonomy S contains simple directed cycles (self loop included). We then use an approach based on the Tarjan algorithm (Tarjan, 1972) to calculate the number

[1] http://www.ebi.ac.uk/chebi/init.do

[2] http://www.google.com/basepages/producttype/taxonomy.en-US.txt

[3] http://www.ise.ncsu.edu/kay/mhetax/index.htm

[4] http://sites.nationalacademies.org/PGA/Resdoc/PGA_044522

of connected components in S. Finally, we compute the number of intermediate nodes as the number of nodes $|V_S| - |L_S|$ where L_S is the set of leaf nodes in S. A leaf node is a node with out-degree = 0.

2.2.3 Comparison against Gold Standard

Previous datasets for evaluating taxonomy extraction (Kozareva et al., 2008) mainly rely on WordNet to gather gold standards from several general knowledge domains, such as animals, plants, and vehicles. The datasets proposed in (Velardi et al., 2013) enrich this experimental setting by including two specialized domains, Virus and Artificial Intelligence, that have low coverage in WordNet. A limitation of these datasets is that currently there is no gold standard taxonomy for these domains, therefore only a manual evaluation is possible. The dataset introduced here, instead, covers four new domains, providing two separate gold standards for each domain: one collected from WordNet, a general purpose resource, and a second one that combines relations from domain-specific resources and from a collaborative resource, Wikipedia, for a higher coverage of the domain. This dataset allows us to investigate how a system performs when taxonomising frequently used terms in comparison with more specialised, rarely used terms.

Given a gold standard taxonomy $G = (V_G, E_G)$, the comparison between a target taxonomy and a gold standard taxonomy is quantified using the following measures:

- common nodes: $|V_S \cap V_G|$
- vertex coverage: $|V_S \cap V_G|/|V_G|$
- number of common edges: $|E_S \cap E_G|$
- edge coverage: $|E_S \cap E_G|/|E_G|$
- ratio of novel edges: $(|E_S| - |E_S \cap E_G|)/|E_G|$
- edge precision: $P = |E_S \cap E_G|/|E_S|$
- edge recall: $R = |E_S \cap E_G|/|E_G|$
- F-score: $F = 2(P * R)/(P + R)$

Additionally, we consider the Cumulative Fowlkes&Mallows (Cumulative F&M) measure (Velardi et al., 2013): the value $B_{S,G}$ between 0.0 and 1.0 which measures level by level how well a target taxonomy S clusters similar nodes compared to a gold standard taxonomy G. $B_{S,G}$ is calculated as follows: let k be the maximum depth of both S and G, and H_{ij} a cut of the hierarchy, where $i \in \{0, ..., k\}$ is the cut level and $j \in \{G, S\}$ selects the clustering of interest. Then, for each cut i, the two hierarchies can be seen as two flat clusterings C_{iS} and C_{iG} 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_{iS} and C_{iG}; n_{00} be the number of object pairs that are in different clusters in both C_{iS} and C_{iG}; n_{10} be the number of object pairs that are in the same cluster in C_{iS} but not in C_{iG}; n_{01} be the number of object pairs that are in the same cluster in C_{iG} but not in C_{iS}.

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

$$B^i_{S,G} = \frac{n^i_{11}}{\sqrt{(n^i_{11} + n^i_{10}) \cdot (n^i_{11} + n^i_{01})}}. \quad (1)$$

And the cumulative Fowlkes&Mallows Measure:

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

2.2.4 Manual quality assessments

The gold standard taxonomies are not complete, therefore it is possible for systems to identify correct relations that are not covered by the gold standard. Normally these relations are considered incorrect using a simple comparison with the gold standard taxonomy. For this reason we manually evaluate a subset of new relations proposed by each system to estimate the number of relations in E_S that do not belong to E_G. A random sample is extracted from all the taxonomies submitted by the participants and then manually annotated to compute the precision P as: $|correctISA|/|sample|$. A total of 100 term pairs were evaluated by three different annotators for each system and each domain, for a total of 800 pairs per system.

The chemical domain is not considered for this evaluation because it requires a considerable amount of domain knowledge and we did not have access to experts in the chemical domain. Two of the authors of this paper independently annotated each sample relation, while the third assessment was done by

a group of five annotators who have a background in Computational Linguistics, with the exception of one annotator who focused on the food domain. Annotators were provided with a list of term pairs organised by domain and were asked if the relation was a correct ISA relation, if the relation and the terms were domain specific, and if the relation was too generic. In our evaluation, a relation is considered correct only if it is a correct hypernym-hyponym relation, if it is relevant for the given domain and not over-generic. Take for example the following edges from the food domain: *(linguine, pasta)* and *(lemon, food)*. Both edges are correct ISA relations and are domain specific, but the second edge is over-generic because lemons are also fruits. The agreement for identifying correct edges is measured using the Fleiss kappa statistic and is overall substantial (Fleiss kappa 0.65). The easiest domain is Food (Fleiss kappa 0.69), followed by Equipment (Fleiss kappa 0.63). Not surprisingly, the Science domain is the most challenging (Fleiss kappa 0.60), as this is a rapidly changing domain and there is in general less consensus about the relations between fields.

3 Submitted runs

Overall, 6 teams participated in the task. Participants were allowed to submit two runs for each of the four domains, one for each type of gold standard, for a total of 8 different runs. Most teams submitted a run for each domain and type of gold standard, with the exception of the LT3 team, which did not submit a system for the Chemical domain and the QASSIT team, which submitted only one run for the WordNet Chemical taxonomy. Next, we will provide a short description of each approach in alphabetical order, discussing corpora collection and the approaches adopted for relation discovery and taxonomy construction.

INRIASAC (supervised) *Corpus:* Wikipedia search using terms; *Relation discovery:* substring inclusion, lexico-syntactic patterns, co-occurrence information based on sentences and documents; *Taxonomy construction:* none.

LT3 (unsupervised) *Corpus:* web corpus constructed using BootCat (Baroni and Bernardini, 2004) using the provided terms as seed terms; *Relation discovery:* lexico-syntactic patterns, morphological structure of compound terms, WordNet lookup (Lefever et al., 2014); *Taxonomy construction:* none.

ntnu (unsupervised) *Corpus:* Wikipedia and WordNet definitions; *Relation discovery:* hypernym extraction from definitions, WordNet lookup, Wikipedia categories, similarity between keywords; *Taxonomy construction:* none.

QASIT (semi-supervised) *Corpus:* Wikipedia, DBpedia; *Relation discovery:* lexico-syntactic patterns, co-occurrence information; *Taxonomy construction:* Learning Pretopological Spaces (LPS) method that learns a Parameterized Space by using an evolutionary strategy.

TALN-UPF (semi-supervised) *Corpus:* Wikipedia definitions retrieved using BabelNet (Navigli and Ponzetto, 2012); *Relation discovery:* based on (Navigli and Velardi, 2010), CRF model trained with the WCL dataset, linguistic rules added to traverse the dependency tree, missing nodes connected to root; *Taxonomy construction:* none.

USAAR (semi-supervised) *Corpus:* Wikipedia documents; *Relation discovery:* lexico-syntactic patterns, co-occurrence information used to construct a vector space model using the word2vec tool;[5] *Taxonomy construction:* none.

4 Results

Table 2 presents the results of the structural analysis (see Section 2.2.2) for all the system outputs and for the two baselines. Only 20 out of 45 submitted taxonomies consist of one weakly connected component (c.c. = 1), and 18 out of 45 are directed acyclic graphs (Cycles=N). Overall, only 10 taxonomies comply with the ideal structural requirements of a taxonomy and are directed acyclic graphs consisting of one connected component. 6 of these were submitted by the only system that addressed the taxonomy construction subtask, QASSIT. Table 3 shows the average edge precision, recall and F-score of the six systems compared to the baselines (see Sections 2.2.3 and 2.2.4). LT3 outperforms the other systems on all the measures. It is worth noting that our string-based baseline (B_2) achieves the

[5] https://code.google.com/p/word2vec/

Table 2: Structural analysis of the submitted taxonomies and of the baseline taxonomies, including the number of: nodes ($|V|$), edges ($|E|$), connected components (c.c.), and intermediate nodes (i.n.).

Combined gold standard taxonomies

		INRIASAC	LT3	ntnu	QASSIT	TALN-UPF	USAAR	B_1	B_2
Chemicals	**|V|**	12432	n.a.	1114	n.a.	17584	13785	17584	10120
	|E|	28444		1563		17606	30392	17583	12672
	c.c.	293		116		**1**	302	**1**	991
	Cycles	Y		**N**		**N**	Y	**N**	**N**
	i.n.	5808		1052		34	13766	1	10117
Equipment	**|V|**	520	260	251	610	612	337	612	248
	|E|	1168	282	247	614	665	548	611	244
	c.c.	6	10	35	**1**	**1**	28	**1**	17
	Cycles	**N**	Y	**N**	**N**	Y	Y	**N**	**N**
	i.n.	164	174	251	70	20	320	1	229
Food	**|V|**	1518	819	834	1550	1549	1118	1549	636
	|E|	4363	1632	1227	1560	1569	2692	1548	627
	c.c.	2	6	27	**1**	**1**	23	**1**	47
	Cycles	Y	**N**	Y	Y	**N**	Y	**N**	**N**
	i.n.	397	159	810	72	18	1105	1	631
Science	**|V|**	417	187	338	453	1280	355	452	232
	|E|	1164	441	386	511	1623	952	451	214
	c.c.	3	8	23	**1**	**1**	14	**1**	28
	Cycles	**N**	Y	**N**	**N**	Y	Y	**N**	**N**
	i.n.	151	88	329	80	422	261	1	207

WordNet gold standard taxonomies

		INRIASAC	LT3	ntnu	QASSIT	TALN-UPF	USAAR	B_1	B_2
Chemicals	**|V|**	1913	n.a.	1475	1351	1347	1173	1351	820
	|E|	4611		1855	1380	1451	3107	1350	808
	c.c.	2		28	**1**	**1**	31	**1**	129
	Cycles	Y		Y	**N**	Y	Y	**N**	**N**
	i.n.	1262		1272	56	63	920	1	819
Equipment	**|V|**	468	462	1081	476	2574	354	475	232
	|E|	1369	1452	1333	490	3370	547	474	188
	c.c.	**1**	**1**	12	**1**	**1**	43	**1**	46
	Cycles	Y	Y	Y	**N**	Y	Y	**N**	**N**
	i.n.	371	142	1036	65	1025	339	1	213
Food	**|V|**	1458	1471	1843	1487	1486	1200	1486	826
	|E|	4238	6913	2760	1539	1548	3465	1485	812
	c.c.	2	**1**	35	**1**	**1**	23	**1**	79
	Cycles	**N**	Y	Y	**N**	**N**	Y	**N**	**N**
	i.n.	478	374	1386	60	53	1189	1	813
Science	**|V|**	366	370	524	371	370	307	370	217
	|E|	1102	1573	681	436	393	892	369	174
	c.c.	**1**	**1**	11	**1**	**1**	8	**1**	48
	Cycles	Y	Y	**N**	**N**	**N**	Y	**N**	**N**
	i.n.	135	114	505	74	25	255	1	208

highest precision, which leads to high F-score, second only to the best system. This is an indication that the test dataset can be improved by removing relations that do not require more sophisticated approaches. The first baseline (B_1) is not competitive, because the gold standard taxonomies are specifically selected to have a rich, deep structure. A large number of novel relations produced by the USAAR system are too generic because they apply a similar strategy. The results of the manual analysis of previously unknown edges are shown in the last line of Table 3. Again, LT3 and INRIASAC systems take the lead. The ntnu system discovers the largest number of novel edges compared to other systems on the WordNet Science taxonomy. In this case, LT3 discovers a larger number of new edges than other participants on Combined taxonomies. In Table 4 we report the Cumulative F&M measure (see Section 2.2.3) for the 45 systems and for the 16 baseline taxonomies. Results are grouped on the basis of the source of the gold standard, that is, combined taxonomies and WordNet taxonomies. LT3 outperforms the other systems on all three submitted WordNet taxonomies by a wide margin (there is no submission for the Chemicals domain), but for the combined taxonomies the INRIASAC system holds the

Table 3: Average Precision, Recall and F-score of ISA relationships across gold standards and Average Precision of novel relations based on human judgement.

	INRIASAC	LT3	ntnu	QASSIT	TALN-UPF	USAAR	B_1	B_2
Comparison against gold standards								
Average Precision	0.1725	0.3612	0.1754	0.1564	0.0720	0.2015	0.0226	**0.5432**
Average Recall	0.4279	**0.6307**	0.2756	0.1589	0.1165	0.3139	0.0212	0.2413
Average F-score	0.2427	**0.3886**	0.2076	0.1575	0.0799	0.2377	0.0219	0.3326
Manual evaluation								
Average Precision	0.4800	**0.5967**	0.4200	0.3533	0.2467	0.1017	-	-

Table 4: Cumulative Fowlkes&Mallows measure for 45 system runs and for 16 baselines.

	INRIASAC	LT3	ntnu	QASSIT	TALN-UPF	USAAR	B_1	B_2
Combined gold standard taxonomies								
Chemicals	**0.2353**	n.a	0.0009	n.a	0.2225	0.00001	0.2281	0.0
Equipment	**0.4905**	0.1137	0.0000	0.4881	0.4482	0.0000	0.3970	0.0012
Food	**0.4522**	0.2163	0.0076	0.3405	0.3267	0.0037	0.3162	0.0007
Science	**0.4706**	0.3303	0.0088	0.5232	0.2202	0.2249	0.4214	0.0108
WordNet gold standard taxonomies								
Chemicals	0.0084	n.a	0.0719	**0.3947**	0.2787	0.2103	0.2683	0.0
Equipment	0.0700	**0.6892**	0.0935	0.3637	0.0901	0.0015	0.2969	0.0007
Food	0.4804	**0.5899**	0.2673	0.3153	0.3091	0.0036	0.2933	0.0022
Science	0.4153	**0.5391**	0.0158	0.2921	0.2126	0.1721	0.1963	0.0016

lead. This difference is explained by the fact that LT3 makes use of a WordNet lookup of hypernym-hyponym relations, which is similar to the method used to collect the WordNet gold standard. More detailed statistics and charts are available on the task website[6]. Finally, in order to obtain an overall rank of the system outputs we first assigned a penalty score (from 1 to 6) for six cue aspects of the evaluation: presence of Cycles, Cumulative F&M measure, number of Intermediate Nodes, F-score from Gold Standard Evaluation, number of Submitted Domains and estimated precision from Manual Evaluation. Then, the total number of penalty points was computed and, following the inverse order of the total penalty scores, we finally ranked the teams (see Table 5).

At the end of the evaluation it emerged that the INRIASAC team had outperformed the other teams in the production of taxonomies for the selected target domains. Although the LT3 team achieved better performance for quantitative approaches (precision, F-score, Cumulative F&M), it was penalised in the final ranking because the constructed taxonomies were generally smaller than the taxonomies produced by INRIASAC, the LT3 team did not submit a taxonomy for Chemicals, and they submitted a larger number of taxonomies with cycles.

Table 5: Overall ranking of submitted systems: INRIASAC (INR), LT3, ntnu, QASSIT (QA), TALN-UPF (TA), USAAR (US).

	INR	**LT3**	**ntnu**	**QA**	**TA**	**US**
Cycles	3	4	2	1	3	4
Cumulative F&M	2	1	6	3	4	5
Intermediate Nodes	2	5	3	6	4	1
Gold Standard Evaluation	2	1	4	5	6	3
Submitted Domains	1	3	1	2	1	1
Manual Evaluation	2	1	4	5	6	3
Total	12	15	20	22	24	17
Final Ranking	1	2	4	5	6	3

5 Discussion

A main limitation of this shared task is that participants were allowed to use the same resources as those used to create the gold standards, and were able to apply simple lookups to retrieve the relations. No recall was computed on the basis of the manual evaluation because of the relatively small number of evaluated relations. A possible solution for this problem would be to use result pooling from all the systems to estimate recall. But this solu-

[6] http://alt.qcri.org/semeval2015/task17/index.php?id=evaluation

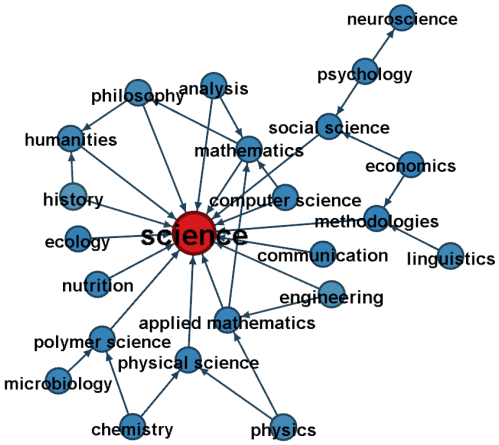

Figure 2: Intermediate nodes of the QASSIT taxonomy on Science.

Figure 3: Intermediate nodes of the gold standard taxonomy on Science.

tion would be more appropriate when there was a larger number of systems. Most participants decided not to address the taxonomy construction subtask, focusing mainly on relation discovery. This could be because the subtask is less well-known and more recently introduced, but also because existing approaches for taxonomy construction are complex and difficult to reimplement. None of the systems was able to address this subtask for the combined Chemicals taxonomy, which is the largest in our dataset. This points to the computational limits of existing algorithms for taxonomy construction. The choice of corpora shows a trend towards using Wikipedia-based corpora instead of web-based corpora (Hovy et al., 2013). Only one participant team relied on web-based corpora. Another lesson that can be drawn from this shared task is that lexico-syntactic patterns, known to have high precision but low recall, can benefit from co-occurrence based approaches, even if these tend to be less reliable. A visualisation of the top levels of the taxonomy constructed by the QASSIT system is presented in Figure 2. The relative size of the nodes within a graph is proportional to the degree of the node. Compared to the gold standard taxonomy for the same domain presented in Figure 3, the QASSIT taxonomy connects a larger number of leaves directly to the Science root, introducing a large number of over-generic relations. There are three times more relations between intermediate nodes and the root node than in the gold standard taxonomy. The QASSIT hierarchy is more shallow than the gold standard, and contains a smaller number of intermediate nodes.

6 Conclusion

This paper provides an overview of the SemEval 2015 task on Taxonomy Extraction. The task aimed to foster research in hierarchical relation extraction from text and taxonomy construction. We constructed and released benchmark datasets for four domains (chemicals, equipment, foods, science). The task attracted 45 submissions from six teams that were automatically evaluated against gold standards collected from WordNet, as well as other well known sources. This evaluation was complemented by a structural analysis of the submitted taxonomies and a manual evaluation of previously unknown edges. Most systems focused on the relation extraction subtask, with the exception of the QASSIT team who addressed the taxonomy construction subtask as well. In future, the datasets can be improved by removing relations that can be identified through string-based inclusion.

Acknowledgements

This work was funded in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289 (INSIGHT) and also by the MultiJEDI ERC Starting Grant No. 259234 (http://multijedi.org/).

References

Marco Baroni and Silvia Bernardini. 2004. Bootcat: Bootstrapping corpora and terms from the web. In *4th Edition of Language Resources and Evaluation Conference (LREC2004)*.

Christiane Fellbaum. 2005. Wordnet and wordnets. In Keith Brown, editor, *Encyclopedia of Language and Linguistics*, pages 665–670, Oxford.

Tiziano Flati, Daniele Vannella, Tommaso Pasini, and Roberto 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 (ACL 2014)*, pages 945–955, Baltimore, Maryland.

Maayan Geffet and Ido Dagan. 2005. The distributional inclusion hypotheses and lexical entailment. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 107–114, Stroudsburg, PA, USA.

Sanda M. Harabagiu, Steven J. Maiorano, and Marius Pasca. 2003. Open-domain textual question answering techniques. *Natural Language Engineering*, 9(3):231–267.

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

Eduard H. Hovy, Roberto Navigli, and Simone Paolo Ponzetto. 2013. Collaboratively built semi-structured content and Artificial Intelligence: The story so far. *Artificial Intelligence*, 194:2–27.

Arthur B. Kahn. 1962. Topological sorting of large networks. *Commun. ACM*, 5(11):558–562.

Zornitsa Kozareva and Eduard Hovy. 2010. A semi-supervised method to learn and construct taxonomies using the web. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1110–1118, Stroudsburg, PA, USA.

Zornitsa Kozareva, Ellen Riloff, and Eduard H Hovy. 2008. Semantic class learning from the web with hyponym pattern linkage graphs. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, volume 8, pages 1048–1056. Citeseer.

Els Lefever, Marjan Van de Kauter, and Véronique Hoste. 2014. Hypoterm: Detection of hypernym relations between domain-specific terms in dutch and english. *Terminology*, 20(2):250–278.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Roberto Navigli and Paola Velardi. 2010. Learning word-class lattices for definition and hypernym extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1318–1327, Uppsala, Sweden, July.

Roberto Navigli, Paola Velardi, and Stefano 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.

Craig Nevill-Manning, Ian Witten, and Gordon W. Paynter. 1999. Lexically-generated subject hierarchies for browsing large collections. *International Journal on Digital Libraries*, 2:111–123.

Mark Sanderson and Bruce Croft. 1999. Deriving concept hierarchies from text. In *In Proceedings of the 22nd annual international ACM SIGIR conference on Research and Development in Information Retrieval*, pages 206–213.

Robert Tarjan. 1972. Depth-first search and linear graph algorithms. *SIAM Journal on Computing*, 1:146–160.

Paola Velardi, Roberto Navigli, Stefano Faralli, and Juana Maria Ruiz-Martinez. 2012. A new method for evaluating automatically learned terminological taxonomies. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey.

Paola Velardi, Stefano Faralli, and Roberto Navigli. 2013. OntoLearn Reloaded: A Graph-Based Algorithm for Taxonomy Induction. *Computational Linguistics*, 39(3):665–707.

Silke Wagner and Dorothea Wagner. 2007. Comparing clusterings an overview. Technical Report 2006-04, Faculty of Informatics, Universität Karlsruhe (TH).

Zhichun Wang, Juanzi Li, and Jie Tang. 2013. Boosting cross-lingual knowledge linking via concept annotation. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, IJCAI '13, pages 2733–2739.