SemEval-2016 Task 13: Taxonomy Extraction Evaluation (TExEval-2)

Georgeta Bordea*, Els Lefever, Paul Buitelaar***
*Insight Centre for Data Analytics
National University of Ireland, Galway
name.surname@insight-centre.org
**LT3, Language and Translation Technology Team,
Ghent University, Belgium
name.surname@ugent.be

Abstract

This paper describes the second edition of the shared task on Taxonomy Extraction Evaluation organised as part of SemEval 2016. This task aims to extract hypernym-hyponym relations between a given list of domain-specific terms and then to construct a domain taxonomy based on them. TExEval-2 introduced a multilingual setting for this task, covering four different languages including English, Dutch, Italian and French from domains as diverse as environment, food and science. A total of 62 runs submitted by 5 different teams were evaluated using structural measures, by comparison with gold standard taxonomies and by manual quality assessment of novel relations.

1 Introduction

Taxonomies are useful tools for content organisation, navigation, and retrieval, providing valuable input for semantically intensive tasks such as question answering (Harabagiu et al., 2003) and textual entailment (Geffet and Dagan, 2005). In general, a hierarchical relation is any asymmetrical relation that indicates subordination between two terms, but in this task we focus on hyponym-hypernym relations. Taxonomy learning from text is a challenging task that can be divided in several subtasks, including term extraction, hypernym identification and taxonomy construction. Existing approaches for hypernym identification from text rely on lexico-syntactic patterns (Hearst, 1992; Lefever et al., 2014), cooccurrence information (Grefenstette, 2015), substring inclusion, or exploit semantic relations provided in textual definitions (Velardi et al., 2013). This stage usually produces a large number of noisy, inconsistent relations, which assign multiple parents to a node and contain cycles. Hence, the third stage of taxonomy learning, taxonomy construction, focuses on the overall structure of the resulting graph and aims to organise terms in a hierarchical structure, more specifically a directed acyclic graph (Velardi et al., 2013; Kozareva and Hovy, 2010).

More recently, the hypernym identification subtask has attracted an increased interest from the distributional semantics community (Santus et al., 2014; Rei and Briscoe, 2014; Roller et al., 2014; Yu et al., 2015), as part of a wider effort to distinguish between different semantic relations which exist between distributional similar words (Weeds et al., 2014; Levy et al., 2015). Although this is a promising direction of research, that addresses some of the limitations of pattern-based approaches, including low coverage of domain-specific terms, most participants in this shared task opted for traditional approaches for hypernym identification, with the exception of one system (Pocostales, 2016).

TexEval-2 is mainly concerned with automatically extracting hierarchical relations from text and subsequent taxonomy construction, therefore we make the assumption that a list of terms is readily available. This simplifies evaluation by providing a common ground for all the systems, but participants are allowed to add additional nodes, i.e. terms, in the hierarchy as they consider appropriate. To avoid the need for term extraction, terms are extracted from existing taxonomies, providing participants with a domain lexicon that has to be organised in a hierarchical structure.

Proceedings of SemEval-2016, pages 1081–1091,
San Diego, California, June 16-17, 2016. ©2016 Association for Computational Linguistics

2 Task Description

The first TExEval shared task (Bordea et al., 2015), organised as part of SemEval 2015, introduced a monolingual dataset that covers terms and hierarchical relations from four domains that were not previously considered for this task. Performance was evaluated across domains, considering common sense knowledge as well as technical domains gathered from WordNet and other well known taxonomies. The second TExEval shared task aimed to extend this experimental setting to a multilingual setting, covering English, French, Italian and Dutch. A main challenge faced by the participants in the first TExEval was that no corpus was provided by the task organisers. We address this issue by providing participants with instructions for downloading and preparing a Wikipedia-based corpus. Depending on the selected approach, a system may or may not require large amounts of text to extract relations between terms, therefore participants are allowed to extend this corpus as they consider appropriate. The task is structured in several subtasks, including monolingual subtasks for hypernym identification and taxonomy construction in English, as well as two corresponding multilingual subtasks that cover Dutch, French and Italian.

3 Dataset Creation

We selected three target domains (i.e. Environment, Food and Science) with three root concepts (i.e. "environment", "food" and "science", respectively). Then, for each domain we considered different sources for gathering gold standard taxonomies, including a multilingual thesaurus, Eurovoc[1], a large lexical database of English, WordNet, and a general purpose resource, the Wikipedia Bitaxonomy (Flati et al., 2014). We also considered other domain-specific resources including "The Google product taxonomy" [2] for Food, and the "Taxonomy of Fields and their Subfields" [3] for Science.

English taxonomies The English gold standard taxonomies are collected from each of the sources described above as follows. Gold standards are gathered from WordNet by selecting concepts and relationships in the hypernym-hyponym hierarchy rooted on the corresponding root concept for each domain. Relations extracted from Wikipedia are combined together with relations extracted from domain-specific resources, to obtain high-coverage domain-specific taxonomies. Hierarchical relations from Eurovoc were used integrally without any modification, but currently Eurovoc covers only the Environment and Science domains. It is worth nothing that the English gold standard taxonomies gathered from WordNet and from combined resources were also used as test data in the previous edition of this shared task (Bordea et al., 2015).

Multilingual taxonomies For the three other languages, the collected English gold standards were manually translated by six linguists (two computational linguists and four master students of the Ghent University Translation, Communication and Interpreting department). In a first step, the English term lists were translated in Excel by one annotator per language. The first annotator was allowed to mark entries that needed to be revised by a second annotator. In addition, the annotators could make remarks in an additional column. Some of the English terms could not be properly translated in the specific domain (e.g. "center" in the food domain) and were left out. In a second step, the translated term lists were used to automatically replace the English terms in the gold standard taxonomies with their corresponding translation.

The translation of English gold standards revealed a number of issues. First of all, some of the translations were near-synonyms in the other language, which eventually lead to cycles in the taxonomy. Examples in Italian are for instance "cibo" (English: food) and "vitto" (English: fare) which are in Italian almost synonymous, whereas their English counterparts have a more distinctive meaning. Another problematic example are the Italian words "condimento" (English: seasoning, sauce, dressing) and "salsa" (English: dressing, sauce), which can be hypernyms of each other, depending on the exact meaning of the word. The translated taxonomies also revealed errors in the original English taxonomy, such as for instance "conserve" is a kind of "confiture", which is incorrect.

[1] Eurovoc: http://eurovoc.europa.eu/drupal/

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

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

Language	Domain	Source	$\lvert V\rvert$	$\lvert E\rvert$	$\#i.i.$	$\#c.c.$	Cycles
English	Environment	Eurovoc	261	261	60	1	no
	Food	Combined	1556	1587	70	1	no
		WordNet	1486	1576	302	1	no
	Science	Combined	453	465	54	1	no
		Eurovoc	125	124	31	1	no
		WordNet	429	452	117	1	no
Dutch	Environment	Eurovoc	267	267	59	1	no
	Food	Combined	1429	1446	66	**3**	no
		WordNet	1299	1340	259	**3**	no
	Science	Combined	445	449	54	1	no
		Eurovoc	125	124	32	1	no
		WordNet	399	399	105	1	no
French	Environment	Eurovoc	267	266	61	1	no
	Food	Combined	1418	1441	64	1	no
		WordNet	1329	1358	263	**2**	no
	Science	Combined	449	451	54	1	no
		Eurovoc	125	124	31	1	no
		WordNet	390	389	101	1	no
Italian	Environment	Eurovoc	267	266	59	1	no
	Food	Combined	1274	1304	60	**3**	no
		WordNet	1277	1332	254	1	**yes**
	Science	Combined	442	444	54	1	no
		Eurovoc	125	124	32	1	no
		WordNet	396	396	105	1	no

Table 1: Structural measures of gold standard taxonomies, including number of vertices ($|V|$), edges ($|E|$), intermediate nodes (#i.i), connected components (#c.c.) and cycles.

Table 1 shows the resulting number of vertices $|V|$ and edges $|E|$ of the produced gold standard taxonomies for each considered domain, source and language. We also report structural information about the number of intermediate nodes (#i.i), the number of connected components (#c.c.) and the number of cycles. Test data for this task consists of six lists of domain-specific terms for each language that were provided to participants as a shared basis to construct the taxonomies. The initial English taxonomies provide connections from the root node to all the other nodes, as they form one connected component. As some of the terms did not have a correspondent in all the other languages, some of the translated taxonomies have several components. This is specifically the case for the food domain that is highly dependent on the language and that shows the largest variation in number of nodes. For example, 127 terms from the Combined English taxonomy for Food could not be translated into Dutch and 279 terms could not be translated into Italian. Additionally, four cycles are erroneously introduced for the WordNet Italian taxonomy for Food, including “cibo”-“vitto”-“cibo” and “piatto principale”-“piatto”-“piatto principale”. Slight differences exist between the Eurovoc taxonomies constructed for different languages as well, and these taxonomies underwent a thorough review process.

4 Evaluation Approach

The construction of taxonomies is a challenging task even for humans but evaluating a taxonomy is not a trivial task either. In this shared task, taxonomies are evaluated through comparison with gold standard relations collected from WordNet and other well known, freely available taxonomies. This is complemented by a manual evaluation of relations that are not covered by the gold standard and through quantitative and qualitative structural analysis of the resulting graph. The evaluation methodology is sim-

Domain	Source	System	$\|V\|$	$\|E\|$	#i.i.	#c.c.	cycles
Environment	Eurovoc	Baseline	123	112	27	17	**no**
		JUNLP	**321**	**463**	123	19	**no**
		TAXI	148	207	50	**1**	**no**
		NUIG-UNLP	312	456	**176**	58	yes
		USAAR	57	47	10	10	**no**
		QASSIT	261	365	88	**1**	**no**
Food	Combined	Baseline	636	627	130	40	**no**
		JUNLP	1802	3015	**581**	48	yes
		TAXI	781	1118	132	**1**	**no**
		NUIG-UNLP	-	-	-	-	-
		USAAR	**3716**	**4347**	323	217	**no**
		QASSIT	-	-	-	-	-
Food	WordNet	Baseline	826	812	205	79	**no**
		JUNLP	**1748**	**3607**	**866**	123	yes
		TAXI	1122	2067	259	**1**	**no**
		NUIG-UNLP	-	-	-	-	-
		USAAR	675	540	146	135	**no**
		QASSIT	-	-	-	-	-
Science	Combined	Baseline	232	214	41	28	**no**
		JUNLP	**602**	1046	255	24	**no**
		TAXI	294	418	73	**1**	**no**
		NUIG-UNLP	595	**1656**	**409**	99	yes
		USAAR	371	312	60	59	**no**
		QASSIT	452	708	58	**1**	yes
Science	Eurovoc	Baseline	50	42	11	9	**no**
		JUNLP	**186**	**342**	**133**	15	yes
		TAXI	100	139	25	**1**	**no**
		NUIG-UNLP	97	218	72	13	yes
		USAAR	37	30	7	7	**no**
		QASSIT	125	164	25	**1**	**no**
Science	WordNet	Baseline	217	174	52	48	**no**
		JUNLP	**424**	690	**304**	90	**no**
		TAXI	290	459	88	**1**	**no**
		NUIG-UNLP	251	**929**	195	9	yes
		USAAR	136	104	32	32	**no**
		QASSIT	370	647	67	**1**	**no**

Table 2: Structural analysis of the submitted taxonomies and the string-based baseline for the monolingual setting

ilar to the approach introduced in the first edition of TExEval, with the main difference that we also report separate overall rankings of the participant systems for each of the subtasks.

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 organizers 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 the string-based baseline described in Section 4.1, 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, favoring systems at the top of the ranked list and penalizing systems at the lower end.

4.1 Baseline

Simple string-based approaches that exploit term compositionality as a main property to hierarchically relate terms are known to be highly effective (Bordea et al., 2015). In this task, we implement the following baseline approach for hypernym extraction and taxonomy construction that is used to benchmark the evaluated systems. The baseline accounts for relations between compound terms such as *(science, network science)* and is implemented as follows:

$$B = (V_B, E_B) \tag{1}$$

where $E_B = \{(a, b)|\ b$ starts with a or ends with a and $|b| > |a|\}$. In this equation a is a term and b is a compound term that includes a as a substring. This baseline approach takes as input only a list of terms and does not require any external corpora or other structured information. It is worth noting that the same approach was applied in the multilingual setting, without any language-specific modification.

4.2 Structural Analysis

In this task, the structural evaluation quantifies the size of a taxonomy under investigation in terms of nodes and edges, evaluating whether the overall graph generated by hypernym-hyponym relations provides connections between the root of the taxonomy and all the other nodes. This is an important property of taxonomies that are used for search because it ensures that all the nodes are findable when exploring the taxonomy from the root. Another structural property of taxonomies is the absence of cycles, which are inconsistent with the semantics of hierarchical relations. Additionally, we highlight the number of nodes located on higher levels of a taxonomy, called intermediate nodes. Finding these nodes is more important than connecting a large number of leaves, as they generate taxonomies with a deeper and richer structure.

Based on these considerations, structural evaluation is performed by computing the cardinality of $|V_S|$ and $|E_S|$. We use an algorithm that finds all the elementary circuits of a simple directed graph (Johnson, 1975) to establish if the taxonomy S contains simple directed cycles (self loop excluded). We then use an approach based on the Tarjan algorithm (Tarjan, 1972) to calculate the number 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, where a leaf node is defined as a node with the outdegree zero.

4.3 Gold Standard Comparison

While initial gold standard datasets for evaluating taxonomy extraction were mainly based on relations extracted from WordNet (Kozareva and Hovy, 2010), more recent work (Velardi et al., 2013) focuses on specialized domains such as artificial intelligence. The dataset introduced in this shared task brings together gold standards collected from WordNet together with gold standards extracted from domain-specific taxonomies and from Wikipedia, a collaborative resource.

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:

- 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), denoted as $B_{S,G}$, and defined as a value 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:

$$B^i_{S,G} = \frac{n^i_{11}}{\sqrt{(n^i_{11} + n^i_{10}) \cdot (n^i_{11} + n^i_{01})}}. \quad (2)$$

And the Cumulative Fowlkes&Mallows Measure:

$$B_{S,G} = \frac{\sum_{i=0}^{k-1} \frac{i+1}{k} B^i_{S,G}}{\sum_{i=0}^{k-1} \frac{i+1}{k}} = \frac{\sum_{i=0}^{k-1} \frac{i+1}{k} B^i_{S,G}}{\frac{k+1}{2}}. \quad (3)$$

4.4 Manual Evaluation

Even the most complete and up to date taxonomy can be extended with additional nodes and relations, therefore it is possible for systems to identify correct relations that are not covered by the gold standard. A problem faced by gold standard evaluation is that these relations are considered incorrect when relying on a direct comparison with the gold standard taxonomy. This is why we additionally evaluate by hand 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. Due to limited resources, we extract only a random sample of novel relations from each submission and manually annotate them to compute precision P as: $|correctISA|/|sample|$. At most 100 relations were evaluated by one annotator for each system, domain, and language for a total of 6200 term pairs. Two different annotators were tasked to evaluate submissions for the monolingual subtask (English) and for the multilingual subtask (Dutch, French, Italian).

The annotators were provided with a list of term pairs organized 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. Overall, a relation is considered correct only if it is considered 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 can be categorized more precisely as fruits.

5 Participants and Results

A total of five teams participated in the shared task, but only two systems participated in the multilingual subtasks. Two of the systems that participated in the monolingual subtask alone did not submit runs for the food domain, which has the largest number of nodes. Overall, 62 system runs were submitted by the five teams, 36 for the multilingual subtasks and 26 for the monolingual subtasks. Next, we provide a short description of each approach starting with the two systems that participated in the multilingual subtasks.

JUNLP The JUNLP system makes use of an external linguistic resource for hypernym identification (Maitra and Das, 2016). This resource is the BabelNet semantic network that connects concepts and named entities in a very large network of semantic relations, called Babel synsets (Navigli and Ponzetto, 2010). To make sure that no relations that were used to construct the gold standards are considered, only relations that mention Wikipedia as a source were selected, discarding relations from all the other sources. Additionally, the system makes use of two string inclusion heuristics. The first heuristic checks if any of the terms provided by the organisers is included as a substring in another term. The second heuristic considers terms that have a considerable overlap, for instance *Chocolate Pudding* and *Vanilla Pudding* although their hypernym (i.e., *Pudding*) is not mentioned in the list of terms. A limitation of this approach is that stopwords are also considered as hypernyms, but this can be easily avoided by using a stopword list.

TAXI The methods for hypernym identification used in the TAXonomy Induction system (TAXI) rely on two sources of evidence: substring matching

Domain	Source	Measure	B	JUNLP	TAXI	NUIG-UNLP	USAAR	QASSIT
Environment	Eurovoc	Fscore	**0.3003**	0.1658	0.2992	0.2008	0.2468	0.1725
		F&M	0.0	0.0814	0.2384	0.0007	0.0007	**0.4349**
Food	Combined	Fscore	0.2665	0.1730	**0.2787**	-	0.0883	-
		F&M	0.0019	**0.2608**	0.2021	-	0.0	-
Food	WordNet	Fscore	0.34	0.2053	0.2932	-	**0.3601**	-
		F&M	0.0022	0.1925	**0.3260**	-	0.0021	-
Science	Combined	Fscore	**0.3947**	0.1906	0.3669	0.1537	0.3063	0.2165
		F&M	0.0163	0.1774	0.3634	0.0090	0.0020	**0.5757**
Science	Eurovoc	Fscore	**0.3133**	0.1931	0.3118	0.1696	0.2468	0.2431
		F&M	0.0056	0.1373	**0.3893**	0.1517	0.0023	**0.3893**
Science	WordNet	Fscore	**0.3834**	0.2487	0.3776	0.2361	0.3058	0.2384
		F&M	0.0016	0.0494	**0.2255**	0.0027	0.0008	**0.2255**

Table 3: Gold standard comparison using Fscore and Cumulative F&M measure for the monolingual setting, where B stands for the string-based baseline

Domain	Source	JUNLP	TAXI	NUIG-UNLP	USAAR	QASSIT
Environment	Eurovoc	0.02	0.11	0.08	**0.22**	0.07
Food	Combined	0.2	0.36	-	**0.73**	-
Food	WordNet	0.18	0.32	-	**0.81**	-
Science	Combined	0.06	0.14	0.09	**0.71**	0.07
Science	Eurovoc	0.02	0.02	0.04	0.0	**0.05**
Science	WordNet	0.06	0.22	0.05	**0.47**	0.22

Table 4: Manual evaluation of 100 (at most) randomly selected novel relations based on precision for English

and Hearst-like patterns (Panchenko et al., 2016). The Hearst patterns for all languages are extracted from Wikipedia and from focused crawls with seed pages that are Wikipedia pages. In addition, for English, several additional corpora are used including GigaWord, ukWaC, a news corpus and the CommonCrawl. For French, Italian and Dutch the method is completely unsupervised and relies on KNN approach. For English, an SVM classifier is trained on the trial data. For all languages the features are the same: substrings and ISA relations extracted with lexico-syntactic patterns. No databases or linguistic resources beyond trial data and raw text corpora mentioned above are used. For the taxonomy construction subtasks, the system makes use of an unsupervised graph pruning approach based on the Tarjan algorithm, connecting the resulting disconnected components to the root of the graph.

NUIG-UNLP The system implements a semi-supervised method that finds hypernym candidates for the provided noun phrases by representing them as distributional vectors. Roughly, this method assumes that hypernyms may be induced by adding a vector offset (Mikolov et al., 2013; Rei and Briscoe, 2014) to the corresponding hyponym representation generated by GloVe over a Wikipedia dump. The vector offset is obtained as the average offset between 200 pairs of hyponym-hypernym in the same vector space selected from trial data.

USAAR This system introduces hypernym endocentricity as a useful property for hypernym identification (Tan, 2016). Often multi-word hyponyms are endocentric constructions which contains a word that fulfills the same function as one part of its word. E.g. an ”apple pie” is essentially a ”pie”. The number of multi-words terms that are endocentric in English is investigated and whether this endocentric property can be used to generate entity links to connect terms in the Wikipedia list of list.

QASSIT A semi-supervised methodology is used for the acquisition of lexical taxonomies based on genetic algorithms (Cleuziou and Moreno, 2016). It is based on the theory of pretopology that offers a powerful formalism to model semantic relations and transforms a list of terms into a structured term space by combining different discriminant criteria.

In particular, rare but accurate pieces of knowledge are used to parameterize the different criteria defining the pretopological term space. Then, a structuring algorithm is used to transform the pretopological space into a lexical taxonomy.

5.1 Monolingual Subtasks (English)

Table 2 presents the results of the structural analysis for English, giving an overview of the structural measures presented in Section 4.2 for each of the submitted runs. The taxonomies constructed by the TAXI and QASSIT systems are the only taxonomies that provide a path from the root to all the other nodes, as the corresponding graphs have a single connected component. All the other submissions have more than ten disconnected components. The string-based baseline is also producing several disconnected components. The TAXI and USAAR systems are the only systems that produce directed acyclic graphs across all the domains. The QASSIT system generates two cycles in the case of the combined taxonomy for Science. Overall, only the TAXI system consistently produces well-structured taxonomies across domains. The systems that output taxonomies with a large number of nodes, edges and intermediate nodes (e.g., JUNLP and NUIG-UNLP) tend to do so at the cost of introducing cycles in the graph.

In Table 3 we summarize the results of the comparison with gold standards in terms of Fscore and the Cumulative F&M measure. The string-based baseline is relatively strong compared to the other systems in terms of Fscore, providing the best results for all the domains with the exception of the food domain. A reason why the baseline is weaker in this domain is that a much larger number of food terms are single-word terms that are not compositional, but in absolute terms the results are still comparative with the results of the best system coming second on the overall ranking. In terms of the Cumulative F&M measure that quantifies structural similarity with the gold standards, the QASSIT system takes the lead for all the domains where a taxonomy was submitted. In the case of the Food domain, it is the TAXI system that achieves the best results for the Combined gold standard and the USAAR system for the WordNet gold standard. The string-based baseline captures only a small part of the structure of the gold standard, as shown by the poor results for the Cumulative F&M measure.

The results of the manual evaluation of a sample of novel relations is presented in Table 4. It is worth noting that not all the systems had at least one hundred novel relations to analyse, therefore in some cases a smaller number of relations was manually evaluated. The USAAR submissions introduce the largest number of correct novel relations, with precision higher than 70% for Food taxonomies and the Science taxonomy gathered from Combined sources. The TAXI system comes second for all the domains with the exception of the Science taxonomy gathered from Eurovoc, where the QASSIT system achieves the best results.

The final ranking of the systems is produced by using a voting approach based on the averaged scores of selected measures that cover the main properties of a well-formed taxonomy. For this shared task all the properties are considered to be equally important, but a weighted approach could also be considered depending on the intended purpose of a taxonomy. These properties include (1) cyclicity, measured in terms of the number of submissions that have cycles; (2) structural similarity with gold standard taxonomies, measured with Cumulative F&M measure; (3) categorization, measured in number of intermediate nodes #i.i. that can be interpreted as taxonomical categories; (4) connectivity, measured in number of connected components #c.c.; (5) overlap of edges with the gold standard taxonomy, measured by Fscore; (6) number of covered domains; (7) precision of novel relations from manual evaluation of sample relations. Table 5 presents the averaged results for each of these measures across domains.

Take for example the best ranked system TAXI, where none of the submitted taxonomies had any cycles, which resulted in an overall score of 0 for cyclicity and a rank 1 in the overall ranking, as this is a desirable feature for a taxonomy. In the case of the structure property, measured by averaging the Cumulative F&M measure over all the submitted taxonomies, the TAXI system achieved the second highest score. This score is below the score achieved by the QASSIT system for the same feature, which brings the TAXI system on the second position in the final ranking for the structure property. Cyclicity

Measure	Baseline	JUNLP	TAXI	NUIG-UNLP	USAAR	QASSIT
Cyclicity	**0**	3	**0**	4	**0**	1
Structure (F&M)	0.01	0.15	0.29	0.04	0.00	**0.4**
Categorisation (#i.i.)	77.67	**377**	104.5	213	96.33	59.5
Connectivity (#c.c.)	36.83	53.17	**1**	44.75	76.67	**1**
GS Fscore	**0.33**	0.20	0.32	0.19	0.26	0.22
Domains	**6**	**6**	**6**	4	**6**	4
Manual Precision	n.a.	0.09	0.20	0.07	**0.49**	0.10

Table 5: Average scores achieved by the systems for the monolingual subtasks.

Subtask	Measure	JUNLP	TAXI	NUIG-UNLP	USAAR	QASSIT
TC	Cyclicity	3	**1**	4	**1**	2
	Structure (F&M)	3	2	4	5	**1**
	Categorisation (#i.i.)	**1**	3	2	4	5
	Connectivity (#c.c.)	3	**1**	2	4	**1**
TC & HI	GS Fscore	4	**1**	5	2	3
	Domains	**1**	**1**	2	**1**	2
	Manual Precision	4	2	5	**1**	3
TC	Ranking	4	**1**	5	3	2
HI		3	**1**	4	**1**	2

Table 6: Overall ranking of the systems for the monolingual subtasks on Taxonomy Construction (TC) and Hypernym Indentification (HI).

and connectivity are the only two properties where low scores are preferable, while for the structure, categorisation, gold standard Fscore, domains, and manual precision higher values are preferred.

The averaged scores shown in Table 5 are directly used to obtain the final ranking of the system for the monolingual subtasks presented in Table 6. The scores used to generate the rankings for the Hypernym Identification (HI) subtask are mainly the last three properties, namely the Fscore with the gold standard, the number of domains, and the precision from manual evaluation. All the seven taxonomical properties described above are used for ranking systems for the Taxonomy Construction (TC) subtask. The TAXI system achieves the best results based on most of these measures, coming third only for the categorization property. This brings the system to the first place both for the Hypernym Identification and the Taxonomy Construction subtasks, in the monolingual setting. There is a tie with the USAAR system, but only for the Hypernym Identification subtask. The second placed system is the QASSIT system, that is ranked on the top three positions for most of the properties with the exception of the categorization property, where it is ranked on the second last place. This is due to the fact that the QASSIT system produces a relatively flat structure, with a smaller number of intermediate nodes.

5.2 Multilingual Subtasks (Dutch, French, Italian)

The results for the multilingual subtasks cover a much smaller number of systems, as only two out of the five participants submitted multilingual taxonomies. The same properties are used for the final rankings of the systems as in the previous section, as can be seen in Table 7. This table shows the average scores of the two systems and of the string-based baseline across domains. Both systems submitted runs for all the domains, therefore in the multilingual subtask the number of domains was not used for ranking the systems. Table 8 presents the final ranking of the systems for the multilingual Taxonomy Construction subtask and the multilingual Hypernym Identification subtask. The TAXI system achieves the best results across all the metrics, with the exception of the categorisation property where JUNLP system introduces a larger number of in-

Measure	Baseline	JUNLP	TAXI
Cyclicity	**0**	**0**	**0**
Structure (F&M)	0.01	0.02	**0.19**
Categorisation (#i.i.)	64.28	**178.22**	64.94
Connectivity (#c.c.)	40.5	34.89	**1**
GS Fscore	**0.31**	0.19	0.28
Manual Precision	n.a.	0.30	**0.63**

Table 7: Average scores of the systems for the multilingual subtasks.

Subtask	Measure	JUNLP	TAXI
TC	Cyclicity	**1**	**1**
	Structure (F&M)	2	**1**
	Categorisation (#i.i.)	**1**	2
	Connectivity (#c.c.)	2	**1**
TC & HI	GS Fscore	2	**1**
	Manual Precision	2	**1**
TC	Ranking	2	**1**
HI		2	**1**

Table 8: Overall ranking of the systems for the multilingual subtasks on Taxonomy Construction (TC) and Hypernym Indentification (HI).

termediate nodes. Again, the string-based baseline achieves the best Fscore results in comparison with the gold standards. The TAXI system achieves the best results for English, with a 12.5% decrease in Fscore for Dutch and French and a 9.4% decrease for Italian. JUNLP performance is more stable across languages, with only a 5% drop in Fscore for Dutch and Italian compared to English, and the same Fscore for French.

6 Conclusion

This paper provides an overview of the SemEval 2016 task on Taxonomy Extraction, that introduced a multilingual dataset for evaluating hypernym extraction and taxonomy construction. The constructed dataset covers three domains including Environment, Food, and Science. The task attracted 62 submissions from five teams that were automatically evaluated against gold standards collected from WordNet, Eurovoc, Wikipedia and other domain-specific resources. We also reported the results of an extensive structural analysis of the submitted taxonomies and a manual evaluation of a sample of edges that are not covered by the gold standards.

The best results were obtained by an approach based on Hearst patterns that makes use of a large web-based corpus including Wikipedia. All the systems could benefit from addressing the taxonomy construction subtask, by paying attention to the overall structure of the taxonomy not just the task of extracting pairs of terms. Compared to the previous edition of TExEval, there are two systems that submitted proper taxonomies compared to just one system last year. In this edition, it is also worthy of mention the introduction of methods that make use of purely distributional approaches. These approaches leave a lot of place for improvement, achieving a competitive recall but lagging behind pattern-based approaches in terms of precision.

A possible improvement of this shared task is to analyse system performance in relation to word polysemy. This could be measured by example based on Wikipedia disambiguation pages or on the number of WordNet senses. It is reasonable to assume that hypernym/hyponym pairs between polysemous words are more difficult to connect without using disambiguation methods to identify the appropriate sense for a domain.

Acknowledgments

This work has been funded in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289 (INSIGHT).

References

Georgeta Bordea, Paul Buitelaar, Stefano Faralli, and Roberto Navigli. 2015. Semeval-2015 task 17: Taxonomy extraction evaluation (TExEval). *SemEval-2015*, 452(465):902.

Guillaume Cleuziou and Jose G. Moreno. 2016. QASSIT at SemEval-2016 Task 13: On the integration of semantic vectors in pretopological spaces for lexical taxonomy acquisition. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, California, June. Association for Computational Linguistics.

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. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Gregory Grefenstette. 2015. INRIASAC: Simple hypernym extraction methods. In *Proceedings of the Ninth International Workshop on Semantic Evaluation (SemEval 2015)*.

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 *Proceedings of the 14th conference on Computational linguistics-Volume 2*, pages 539–545. Association for Computational Linguistics.

Donald B Johnson. 1975. Finding all the elementary circuits of a directed graph. *SIAM Journal on Computing*, 4(1):77–84.

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. Association for Computational Linguistics.

Els Lefever, Marjan Van de Kauter, and Veronique Hoste. 2014. HypoTerm: Detection of hypernym relations between domain-specific terms in Dutch and English. *Terminology*, 20(2):250–278.

Omer Levy, Steffen Remus, Chris Biemann, Ido Dagan, and Israel Ramat-Gan. 2015. Do supervised distributional methods really learn lexical inference relations. *Proceedings of NAACL, Denver, CO*.

Promita Maitra and Dipankar Das. 2016. JUNLP at SemEval-2016 Task 13: A language independent approach for hypernym identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, California, June. Association for Computational Linguistics.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pages 746–751.

Roberto Navigli and Simone Paolo Ponzetto. 2010. Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 216–225. Association for Computational Linguistics.

Alexander Panchenko, Stefano Faralli, Eugen Ruppert, Steffen Remus, Hubert Naets, Cedrick Fairon, Simone Paolo Ponzetto, and Chris Biemann. 2016. TAXI at SemEval-2016 Task13: a taxonomy induction method based on lexico-syntactic patterns, substrings and focused crawling. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*. Association for Computational Linguistics.

Joel Pocostales. 2016. NUIG-UNLP at SemEval-2016 Task 13: A simple word embedding-based approach for taxonomy extraction. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, San Diego, California, June. Association for Computational Linguistics.

Marek Rei and Ted Briscoe. 2014. Looking for hyponyms in vector space. In *CoNLL*, pages 68–77.

Stephen Roller, Katrin Erk, and Gemma Boleda. 2014. Inclusive yet selective: Supervised distributional hypernymy detection. In *COLING*, pages 1025–1036.

Enrico Santus, Alessandro Lenci, Qin Lu, and Sabine Schulte Im Walde. 2014. Chasing hypernyms in vector spaces with entropy. In *EACL*, pages 38–42.

Liling Tan. 2016. USAAR at SemEval-2016 Task 13: Hyponym endocentricity. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, California, June. Association for Computational Linguistics.

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

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).

Julie Weeds, Daoud Clarke, Jeremy Reffin, David Weir, and Bill Keller. 2014. Learning to distinguish hypernyms and co-hyponyms. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2249–2259.

Zheng Yu, Haixun Wang, Xuemin Lin, and Min Wang. 2015. Learning term embeddings for hypernymy identification. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, pages 1390–1397. AAAI Press.