SemEval-2018 Task 9: Hypernym Discovery

Jose Camacho-Collados[♣] **Claudio Delli Bovi**[♡] **Luis Espinosa-Anke**[♣]
Sergio Oramas[♢] **Tommaso Pasini**[♡] **Enrico Santus**[♥]
Vered Shwartz[♠] **Roberto Navigli**[♡] **Horacio Saggion**[♢]
[♣]School of Computer Science and Informatics, Cardiff University, United Kingdom
[♡] Computer Science Department, Sapienza University of Rome, Italy
[♢] Pompeu Fabra University, Barcelona, Spain [♥] MIT, United States
[♠] Bar-Ilan University, Ramat Gan, Israel
[♣]{camachocolladosj,espinosa-ankel}@cardiff.ac.uk,
[♡]{dellibovi,pasini,navigli}@di.uniroma1.it,
[♢]{name.surname}@upf.edu, [♥]esantus@mit.edu, [♠]vered1986@gmail.com

Abstract

This paper describes the SemEval 2018 Shared Task on Hypernym Discovery. We put forward this task as a complementary benchmark for modeling hypernymy, a problem which has traditionally been cast as a binary classification task, taking a pair of candidate words as input. Instead, our reformulated task is defined as follows: given an input term, retrieve (or discover) its suitable hypernyms from a target corpus. We proposed five different subtasks covering three languages (English, Spanish, and Italian), and two specific domains of knowledge in English (Medical and Music). Participants were allowed to compete in any or all of the subtasks. Overall, a total of 11 teams participated, with a total of 39 different systems submitted through all subtasks. Data, results and further information about the task can be found at https://competitions.codalab.org/competitions/17119.

1 Introduction

Hypernymy, i.e. the capability to relate generic terms or classes to their specific instances, lies at the core of human cognition. It is not surprising, therefore, that identifying hypernymic (*is-a*) relations has been pursued in NLP for more than two decades (Shwartz et al., 2016): indeed, successfully identifying this lexical relation substantially improves Question Answering applications (Prager et al., 2008; Yahya et al., 2013), Textual Entailment and Semantic Search systems (Hoffart et al., 2014; Roller et al., 2014; Roller and Erk, 2016). In addition, hypernymic relations are the backbone of almost every ontology, semantic network and taxonomy (Yu et al., 2015), which are in turn useful resources for downstream tasks such as web retrieval, website navigation or records management (Bordea et al., 2015).

Generally, evaluation benchmarks for modeling hypernymy have been designed such that in most cases they are reduced to binary classification (Baroni and Lenci, 2011; Snow et al., 2004; Boleda et al., 2017; Vyas and Carpuat, 2017), where a system has to decide whether a hypernymic relation holds between a given candidate pair of terms. Criticisms to this experimental setting point out that supervised systems tend to benefit from the inherent modeling of the datasets in the hypernym detection task, leading to lexical memorization phenomena (Levy et al., 2015; Santus et al., 2016a; Shwartz et al., 2017). In this respect, recent work has attempted to alleviate this issue by including a graded scale for evaluating the degree of hypernymy on a given pair (Vulić et al., 2017).

Crucially, Espinosa-Anke et al. (2016) proposed to frame the problem as *Hypernym Discovery*, i.e. given the search space of a domain's vocabulary, and given an input term, discover its best (list of) candidate hypernyms. This formulation addresses one of the main drawbacks of the evaluation criterion described above, and better frames the evaluated systems within downstream real-world applications (Camacho-Collados, 2017). In fact, lessons learned from these studies have motivated the construction of a full-fledged benchmarking dataset for the shared task we present here, which covers multiple languages and knowledge domains. The main goal of this task is that of complementing current research in hypernymy modeling with this novel discovery setting.

Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018), pages 712–724
New Orleans, Louisiana, June 5–6, 2018.

	Term	Hypernym(s)	Source
1A: English	sorrow	sadness, unhappiness	WordNet
1B: Italian	Nina Simone	musicista, pianista, persona	MultiWibi
1C: Spanish	guacamole	salsa para mojar, salsa, alimento	Wikidata (via BabelNet)
2A: Medical	pulmonary embolism	pulmonary artery finding, trunk arterial embolus, embolism	SnomedCT
2B: Music	Green Day	artist, rock band, band	MusicBrainz

Table 1: Some example terms and hypernyms extracted from different sources (see Section 4.1.4), for each of the subtasks and languages considered in the task.

2 Related Work

Traditionally, identifying hypernymic relations from text corpora has been addressed with two main approaches: pattern-based and distributional (Wang et al., 2017). Pattern-based (path-based) methods, which provide higher precision at the price of lower coverage, exploit the co-occurrence of a hyponym and its hypernym in a textual corpus (Hearst, 1992; Navigli and Velardi, 2010; Boella and Di Caro, 2013; Flati et al., 2016; Gupta et al., 2016; Pavlick and Pasca, 2017). Conversely, distributional models rely on a distributional representation for each observed word, and are capable of identifying hypernymic relations between concepts even when they do not co-occur explicitly in text. Earlier work on hypernym modeling was unsupervised, and leveraged various interpretations of the distributional hypothesis.[1] Most of the recent work on the subject is however supervised, and in the main based on using word embeddings as input for classification or prediction (e.g Baroni et al., 2012; Santus et al., 2014; Fu et al., 2014; Weeds et al., 2014; Espinosa-Anke et al., 2016; Sanchez Carmona and Riedel, 2017; Nguyen et al., 2017). As shown by Shwartz et al. (2016), pattern-based and distributional evidences can be effectively combined within a neural architecture. In this shared task we have actually received systems of both natures, including a combination of pattern-based and distributional cues, similar to the one mentioned above, which also proved to be highly effective (see Section 5).

3 Task Description

We define Hypernym Discovery operatively as the task of finding and extracting the appropriate hypernym(s) for a target input term. As input for the task, together with the target term,[2] a large textual corpus (*source corpus* henceforth) is provided, and participating systems are intended to exploit this large source of textual data to retrieve (i.e. "discover") as many suitable hypernyms as possible for the target term. A different source corpus, as well as the corresponding vocabulary, is specified for each subtask and language (cf. Section 4) in order to set a level playing field for competing systems, and constrain their search space.

For each input term (or *hyponym*) the expected output is a ranked list of *candidate hypernyms* (up to 15) drawn from the provided vocabulary. Some example input-output pairs (i.e. terms and corresponding hypernym lists) are shown in Table 1 for each subtask and language. Table 1 also reports the sources of hypernymy information beside each pair, which vary depending on the subtask and language, as detailed in Section 4.1.4.

The structure of our Hypernym Discovery task consists of five independent but related subtasks, split into two larger groups: *general-purpose* hypernym discovery and *domain-specific* hypernym discovery. Participants were allowed to submit systems for any individual subtask. Along with a specific source corpus and vocabulary, each subtask features its specific training and testing data, consisting of input terms and corresponding gold hypernym lists, obtained as described throughout Section 4.

General-Purpose Hypernym Discovery consists in discovering hypernyms in a large corpus of general-purpose textual data, gathered from different and heterogeneous sources. A system operating in this setting requires the flexibility to provide hypernyms for terms in a wide range of domains. In this shared task we consider three different lan-

[1] See Shwartz et al. (2017) for a detailed review on unsupervised distributional hypernymy detection.

[2] A valid input term is any word or multi-word expression drawn from the predefined vocabulary (cf. Section 4.1.2) up to trigrams.

guages for general-purpose hypernym discovery:

- **English** (**subtask 1A**), with a gold standard of 3,000 labeled terms;
- **Italian** (**subtask 1B**) and **Spanish** (**subtask 1C**), each with a gold standard of 2,000 labeled terms;

All the gold standards provide a balanced set of input terms, with different degrees of frequency and for different domains. The corresponding gold hypernyms have been extracted from multiple resources and manually validated (cf. Sections 4.1.4-4.1.5). Training and testing data are split evenly (50% training - 50% testing).

Domain-Specific Hypernym Discovery deals with the same problem, but constrains it to a specific domain of knowledge. As a consequence, in this case participants test their systems (which might be general or specifically tailored to the target domain) in a much more focused and reduced environment. In this shared task we focus on English and consider two different domains of knowledge:

- **Medical** (**subtask 2A**), with a gold standard of 1,000 labeled terms;
- **Music** (**subtask 2B**), also with a gold standard of 1,000 labeled terms;

As in the previous subtask, we provide a balanced set of terms and gold hypernyms, with different degrees of frequency and for different subdomains. Again, training and testing data are split evenly (50% training - 50% testing).

Subclass vs. Instance. Although many hypernym detection approaches tend to overlook this distinction, it is customary to consider two different varieties of the "is-a" relation: a `subclass-of` variety (e.g. a dog *is a* mammal), and an `instance-of` variety (e.g. Rome *is a* city).[3] From a practical standpoint, the former occurs between two concepts, while the latter connects a named entity with a concept. We make this distinction explicit in our shared task by hand-labeling each input term as either a concept or a named entity. This strategy serves a double purpose: on one hand, it helps reducing lexical ambiguity, and narrowing the search space of potential hypernyms even further;[4] on the other hand, it enables participants to study and develop models specifically tailored to one of the two varieties, and possibly submit them separately. In this respect, Boleda et al. (2017) has indeed shown how systems tend to perform differently on these two kinds of hypernymy relation.

[3] In fact, WordNet encodes `hypernym` and `instance` as two separate semantic relations. Instances are always leaf (terminal) nodes in their hierarchies.

4 Task Data

In this section we present the data collection process carried out for each source corpus and gold standard featured in the task (Section 4.1). We then summarize and provide some global statistics on all these datasets (Section 4.2).

4.1 Data Collection Process

The process of collecting data for each subtask and language comprised five successive steps: compilation of the source corpus (Section 4.1.1), creation of the vocabulary (Section 4.1.2), collection and selection of the input terms (Section 4.1.3), extraction of the gold hypernyms (Section 4.1.4), and final filtering and validation of such hypernyms (Section 4.1.5).

4.1.1 Corpus Compilation

First, we selected and compiled a source corpus for each dataset, which was also considered in the vocabulary creation step (Section 4.1.2). Naturally, we considered three corpora as general and as large as possible for the general-purpose track, whereas for the domain-specific datasets we opted for more targeted and specific text collections.

General-purpose corpora. As source corpus for the English subtask (1A) we used the 3-billion-word UMBC corpus[5] (Han et al., 2013), which is a resource composed of paragraphs extracted from the web as part of the Stanford WebBase Project[6] (Hirai et al. 2000). The UMBC corpus is considerably large and contains information from many and diverse domains. This corpus presents additional challenges and different

[4] As an example, the term *apple* could either refer to a fruit (if labeled as concept) or to a company (if labeled as named entity).

[5] `http://ebiquity.umbc.edu/blogger/2013/05/01/umbc-webbase-corpus-of-3b-english-words/`

[6] `http://dbpubs.stanford.edu:8091/~testbed/doc2/WebBase/`

sources of information with respect to the corpora used in previous tasks, such as Wikipedia in the SemEval 2016 task on taxonomy extraction (Bordea et al., 2016). In fact, the encyclopedic nature of Wikipedia has been exploited in a wide variety of works (Ponzetto and Strube, 2007; Flati et al., 2016; Gupta et al., 2016), and differs substantially from the web-based corpus we put forward here. As source corpus for the Italian subtask (1B) we instead used the 1.3-billion-word itWac corpus[7] (Baroni et al., 2009), extracted from different sources of the web within the `.it` domain. Finally, as source corpus for the Spanish subtask (1C) we considered the 1.8-billion-word Spanish corpus[8] (Cardellino, 2016), which also contains heterogeneous documents from different sources.

Domain-specific corpora. As source corpus for the medical domain (subtask 2A) we provided a combination of texts drawn from the MEDLINE[9] (Medical Literature Analysis and Retrieval System) repository, which contains academic documents such as scientific publications and paper abstracts. This corpus contains 130 million words. As regards the music domain (subtask 2B), instead, the source corpus we compiled is a concatenation of several music-specific corpora, i.e. music biographies from `Last.fm` contained in ELMD 2.0 (Oramas et al., 2016), articles from the music branch of Wikipedia, and a corpus of album customer reviews from Amazon (Oramas et al., 2017). The resulting corpus reaches 100 million words in total.

4.1.2 Vocabulary Creation

With the aim of simplifying the task for participants by providing a unified hypernym search space, we built a series of vocabulary files including all the possible hypernyms on each dataset. Each vocabulary was constructed by considering all the words occurring at least N times across the source corpus of the corresponding subtask. We set N to five and three in the general-purpose and domain-specific subtasks, respectively. We also included bigrams and trigrams, by considering all the instances present in any of the resources that we leveraged as part of the hypernym extraction process (see Section 4.1.4), provided that they also surpassed the corresponding frequency thresholds.

In order to reduce the high granularity of some hypernymy relations (for example, *dog* is an *entity*) we created an additional blacklist of very general terms not considered in the vocabulary files. This list was obtained semi-automatically. We first extracted the most common hypernyms from the lexical sources we used for creating the datasets. Then, we filtered the resulting blacklist by removing manually a number of suitable hypernyms that, despite being general, provided useful information worthy to be taken into account (e.g. *animal*).

4.1.3 Term Collection

After compiling a source corpus and a corresponding vocabulary, we selected a suitable collection of input terms (i.e. hyponyms) to construct the gold standard for each subtask. Term selection was based on three key constraints. First, as in vocabulary creation step (Section 4.1.2), input terms were required to occur five and three times in the general-purpose and domain-specific datasets, respectively. Second, only terms up to trigrams were considered. Finally, we only allowed terms with at least one extracted hypernym (see Section 4.1.4) present in the corresponding vocabulary file.

We carried out the term collection process with a semi-automatic two-pass procedure, which we applied to the source corpus of each subtask. First, candidate terms were extracted automatically from the source corpus, taking into account frequency, type (i.e. concept and entity) and knowledge domain[10] in order to produce a list as balanced and representative as possible. After a preliminary list of input terms was obtained, we carried out an extensive validation and refinement step by manually normalizing each item (e.g. changing plurals to singulars, capitalizing named entities and lowercasing concepts), and by pruning all the terms that appeared too vague or general, as well as terms with mis-attributed domains.

4.1.4 Automatic Hypernym Extraction

Once the terms were collected we proceeded to extract a set of candidate hypernyms from a number of heterogeneous taxonomies. We drew taxonomic information from the following lexical resources: WordNet (Miller, 1995), Wikidata

[7] http://wacky.sslmit.unibo.it/doku.php?id=corpora

[8] http://crscardellino.me/SBWCE/

[9] https://www.nlm.nih.gov/databases/download/pubmed_medline.html

[10] We leveraged the domains from the Wikipedia featured articles pages available in BabelDomains (Camacho-Collados and Navigli, 2017).

(Vrandečić and Krötzsch, 2014), MultiWiBi (Flati et al., 2016), and Yago (Suchanek et al., 2007). In order to be able to use seamlessly all hypernymy information for languages other than English, we exploited the inter-resource mappings provided by BabelNet (Navigli and Ponzetto, 2012).[11] For the domain-specific datasets we additionally used SnomedCT (Spackman et al., 1997) and MusicBrainz (Swartz, 2002) for the medical and music datasets, respectively.

The hypernym extraction process was carried out as follows: given a term (hyponym), we first retrieved all the BabelNet synsets which included the given term as lexicalization; then, starting from that synset, we iteratively visited the father nodes across all the reference taxonomies up to five levels[12] and selected all the lexicalizations of the traversed synsets (i.e. concepts) as given by BabelNet, provided that they appeared in the corresponding vocabulary files (see Section 4.1.2).

4.1.5 Hypernym Validation

Starting from the candidate gold hypernyms extracted in the previous step, we carried out a validation step using human annotators. We leveraged crowdsourcing for the English data in subtask 1A (which featured the largest dataset), and then expert verification in all subtasks (including English).

Crowdsourcing. We validated the English gold standard (both training and test set) by using crowdsourcing workers from Amazon Mechanical Turk. To ensure the quality of workers, we required workers to have answered at least 500 prior HITs with an approval rate of at least 95%, and applied a qualification test. For each target term, we showed the workers multiple candidate hypernyms, extracted in the previous step (Section 4.1.4), and asked them to select all the correct hypernyms. We also added 20% of random false candidates to prevent bias towards a positive answer. Finally, we assigned each HIT to 3 workers and determined the gold label with majority voting. The resulting annotations yielded an inter-annotator agreement of 73%.

[11] Yago is the only resource which is not mapped to BabelNet. For the mapping we simply relied on the WordNet and Wikipedia identifiers provided in Yago.

[12] We decided to consider only five levels for two reasons: first, to avoid very general hypernyms; and second, to avoid errors which would propagate to other levels and make the validation task much harder. To this aim, five levels seemed to provide a fine balance between precision and recall.

	1A	1B	1C	2A	2B
Trial	50	25	25	15	15
Training	1,500	1,000	1,000	500	500
Test	1,500	1,000	1,000	500	500

Table 2: Number of terms (hyponyms) for each dataset in trial, training and test sets.

Expert verification. Expert verification comprised two steps. First, all the extracted data was verified by an expert human annotator. In this first step, the annotator was focused on removing the incorrect hypernyms, or normalizing them if required (e.g. plural to singular). This first verification was performed in all dataset except English, which underwent the crowsourcing validation explained earlier. Then, all datasets (including the English one) were again verified by other experts. However, in this case the annotators were given different guidelines: in particular, they were asked to fix clear hypernym errors (which may have been missed in the previous step) and to add obvious hypernyms which they found to be missing.

4.2 Statistics

Table 2 shows the number of input terms in each dataset. The dataset was split equally in training and testing, while the trial data provided a fewer examples and could also be used as development set. English (subtask 1A) was the largest dataset with 1,500 terms (hyponyms) and for training and other 1,500 for testing. Then, for the Italian (subtask 1B) and Spanish (subtask 1C) datasets, 2,000 terms were given overall between training and testing. Finally, both domain-specific datasets (i.e. medical, subtask 2A, and music, subtask 2B) contained half of this quantity, with 1,000 terms each.

Note that each term may be associated with one or (in most cases) more than one hypernym. Therefore, counting all the term-hypernym pairs per dataset, as it is done in hypernymy detection datasets, would provide much larger figures. As an example, the number of term-hypernym pairs in the test gold standard is 7,048 for English, 4,770 for Italian, 6,070 for Spanish, 4,116 for the medical dataset, and 5,233 for the music dataset.

5 Evaluation

Parting ways from the classic precision-recall-F_1 metrics used so far in hypernym detection/extraction, we decided to evaluate this shared

task as a soft ranking problem. Systems were evaluated over the top 15 (at most) hypernyms retrieved for each input term, which let us assess their performance through Information Retrieval metrics. Let us briefly introduce each of them.

Mean Average Precision (MAP). We use MAP as the main evaluation metric of this task. Intuitively, this metric should give a fine estimate on the capability of a system to retrieve a sizable number of hypernyms from textual data, as well as considering the precision of each of them. Formally:

$$\text{MAP} = \frac{1}{|Q|} \sum_{q \in Q} \text{AP}(q)$$

where Q is a sample of experiment runs, AP($\cdot$) refers to *average precision*, i.e. an average of the correctness of each individual obtained hypernym from the search space.

Mean Reciprocal Rank (MRR). MRR rewards the position of the first correct result in a ranked list of outcomes, and is defined as:

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i}$$

where $rank_i$ refers to the rank position of the *first* relevant outcome for the *i*th run. While its main field of application is Information Retrieval, it has also been used in NLP tasks such as collocation recognition (Wu et al., 2010; Rodríguez-Fernández et al., 2016).

In addition to the above, we also provide results according to **P@*k***, i.e. the number of correctly retrieved hypernyms at different cut-off thresholds, specifically $k \in \{1, 3, 5, 15\}$.[13]

5.1 Baselines

We compared the participating systems with both supervised and unsupervised baselines for each subtask, inspired by recent work on hypernym detection and discovery. In this section we briefly describe each of them.

5.1.1 Supervised Baselines

We first used a naïve *most frequent hypernym* (**MFH**) baseline, which simply returns, for each input term, the 15 most frequent hypernyms found in the training data. As a less naïve baseline, we also trained a *transformation matrix* (Mikolov et al., 2013; Fu et al., 2014), using the same optimization described by Espinosa-Anke et al. (2016). For this baseline the hypernyms in the vocabulary which are among the fifteen closest vectors by applying the transformation matrix are retrieved. However, unlike in the original implementation, in this case we did not perform any a priori domain clustering of the embeddings space, and thus used the same matrix for all input terms.[14] This second supervised baseline is referred to as **vTE** (*vanilla Taxoembed*).

5.1.2 Unsupervised Baselines

We developed an unsupervised baseline by reducing hypernymy discovery to hypernymy detection. We generated a list of candidate hypernyms for each target word, and then employed unsupervised hypernymy detection measures to decide whether a hypernymy relation holds. We used the open-source code by Shwartz et al. (2017).[15]

Our baseline starts by creating a distributional semantic model (DSM) for each domain/language (English, Spanish, Italian, Music and Medical). We used a non-directional window of size 5 as context type, and PPMI as feature weighting. Similarly to the hyponym selection step (Section 4.1.3), all the terms with frequency of at least 3 occurrences in the source corpus are considered as valid targets. For the context words, instead, we required a minimum of 100 occurrences, as in Shwartz et al. (2017). To generate candidates, we took the 50 most similar terms for each target word via cosine similarity in the DSM.

We chose the hypernym detection measures as representative algorithms from each "family" of unsupervised measures: **APSyn** (Santus et al., 2016b) as similarity measure, **balAPInc** (Kotlerman et al., 2010) as measure based on the distributional inclusion hypothesis, and **SLQS** (Santus et al., 2014) as measure based on informativeness.[16] Finally, we tuned the thresholds for the above measures by maximizing the average of the performance metrics on the training set, separately for each subtask and measure.

[13] Although only P@5 is displayed in the tables due to lack of space, the other thresholds were used in the official evaluation as well.

[14] We used the open-source code available at https://bitbucket.org/luisespinosa/taxoembed

[15] https://github.com/vered1986/UnsupervisedHypernymy

[16] Following the conclusions from Shwartz et al. (2017), we set the hyper-parameters to: SLQS: median, PLMI, $N = 100$ and APSyn: $N = 500$.

5.2 Participant Systems

Table 3 shows a summary of all participant systems, displaying their main features with respect to supervison and external resources used, if any.

5.3 Results

A summary of the results is provided in tables 3 to 7, respectively describing results for English, Italian, Spanish and Music and Medical domains. Almost all systems performed better than the unsupervised baselines, while the supervised ones showed to be more challenging, with few systems outperforming them. For English, Music and Medical domains, **CRIM** (Bernier-Colborne and Barriere, 2018) obtained the best results, with a large margin on the other systems and baselines. This system is based on learning a projection between hyponym-hypernym pairs in terms of their corresponding embeddings, and combines this module with an unsupervised system which uses Hearst-style patterns. Moreover, in Italian, the best system was **300-sparsans_r1** (Berend et al., 2018), a logistic regression model informed mostly with information coming from word embeddings; whereas for Spanish, the best performing team was **NLP_HZ** (Qiu et al., 2018), who approached the task with a nearest neighbors algorithm trained with the provided training data.

From the summary tables we can also appreciate the difference in performance of the systems on concepts and entities. Such difference is due to several factors, including the quantity and type of hypernyms that needed to be identified for the two subclasses. Except for the Music domain, systems tended to perform better with entities than with concepts. This is probably due to the fact that entities contain many hypernyms which appear often (e.g. *person*, *company*), which in principle favor the inherent lexical memorization (Levy et al., 2015) of supervised systems. Hence, as expected, systems performed better in the specialized domains (i.e. medical and music) than in the general-domain dataset (34.05% and 40.97% MAP performance by the best systems in the medical and music domains, respectively, compared to the 19.78% result of the best system in the English dataset).

Finally, the results also show the clear superiority of supervised systems over unsupervised approaches in all languages and domains. As far as fully unsupervised systems are concerned, they achieved a diverse degree of success. While in general they were outperformed by supervised systems, in some cases their performance came close, especially for concepts. For instance, the **ADAPT** (Maldonado and Klubika, 2018) system, which is based on a simple similarity measure applied to word embeddings, achieved a very decent 8.13 MAP percentage performance on the medical dataset, using neither supervision nor external resources. Supervised systems produced a larger gap for entities, probably due, as mentioned above, to the lower diversity of possible hypernyms.

Cross-evaluation. In addition to the normal setting on which supervised systems trained their system on the same dataset training data, we ask participants to train systems on the English general-purpose data and trained on the domain-specific datasets. This experiment could enable us to test how a system could perform on a particular dataset when training data is not available. A few teams provided results on this setting and the results showed that even though trained on general data, they are still competitive with respect to other approaches. In fact, they tend to equally outperform unsupervised systems and in the medical dataset, for example, CRIM trained on the general English corpora outperformed all remaining participant systems trained on the medical training data.

6 Analysis

Inspired by previous tasks in taxonomy learning (Bordea et al., 2015), we sampled for each system 50 incorrect hypernyms (25 entities, 25 concepts) which were retrieved as first choice, and manually assessed their correctness. This evaluation of false positives is intended to account for the inevitable scenario in which not all possible correct hypernyms according to human judgement were included in the gold standard. The results in false positives were measured by accuracy (i.e. percentage of correct false positives on the given sample) and are displayed in Tables 4-8 under *FPs*.

In general, we observe that the systems' performances in this false positives experiment are correlated with the figures they obtained with the other automatic evaluation measures. Nonetheless, according to this false positives evaluation, most systems (both supervised and unsupervised) were able to retrieve some hypernyms which were not present in the gold standard. This result is encouraging, as not only hypernym discovery sys-

	Team Name	Reference	Supervision	External Resources
Systems	CRIM	(Bernier-Colborne and Barriere, 2018)	✓	-
	MSCG-SANITY	-	✓	Microsoft Concept Graph
	NLP_HZ	(Qiu et al., 2018)	✓	-
	300-sparsans	(Berend et al., 2018)	✓	-
	SJTU BCMI	(Zhang et al., 2018)	✓	-
	UMDuluth	(Hassan et al., 2018)	✓	-
	ADAPT	(Maldonado and Klubika, 2018)		-
	Apollo	(Onofrei et al., 2018)		-
	EXPR	(Issa Alaa Aldine et al., 2018)		-
	Team 13	-		-
	Anu	-		WordNet
Baselines	vanillaTaxoEmbed	(Espinosa-Anke et al., 2016)	✓	-
	MFH	-	✓	-
	APSyn	(Shwartz et al., 2017)		-
	balAPInc	(Shwartz et al., 2017)		-
	SLQS	(Shwartz et al., 2017)		-

Table 3: Summary of participating systems and baselines, along with their main features (i.e. with or without supervision, and usage of external resources).

1A: English												
	Concepts				**Entities**				**All**			
	MAP	MRR	P@5	FPs	MAP	MRR	P@5	FPs	MAP	MRR	P@5	FPs
CRIM_r1	**16.08**	**30.04**	**15.41**	20	**29.21**	**51.82**	**27.74**	24	**19.78**	**36.10**	**19.03**	22
CRIM_r2	15.49	29.29	14.97	24	28.63	50.55	27.65	20	19.54	35.94	18.74	22
MSCG-SANITY_r1	9.36	18.9	9.38	28	17.72	38.85	16.91	20	11.83	24.79	11.60	24
vTE*	6.99	16.05	6.55	**36**	19.22	42.39	17.92	12	10.60	23.83	9.91	24
MSCG-SANITY_r2	8.66	17.24	8.76	24	12.49	28.20	12.09	**40**	9.80	20.48	9.74	**32**
NLP_HZ	7.17	13.13	7.11	24	14.61	27.21	14.14	20	9.37	17.29	9.19	22
300-sparsans_r1	6.41	13.92	6.33	24	15.02	32.61	14.10	16	8.95	19.44	8.63	20
MFH*	4.73	12.48	4.13	0	18.42	42.65	16.59	16	8.77	21.39	7.81	8
300-sparsans_r2	5.97	12.72	5.73	20	14.78	30.62	14.21	20	8.58	18.00	8.23	20
SJTU BCMI	3.29	5.68	3.57	0	11.70	22.19	11.67	12	5.77	10.56	5.96	6
Team 13	3.70	7.92	3.66	12	0.52	1.65	0.46	20	2.77	6.07	2.72	16
Apollo_r2	2.72	6.05	2.76	16	2.60	5.91	2.51	20	2.68	6.01	2.69	18
Apollo_r1	1.36	3.28	1.34	16	1.48	4.05	1.31	16	1.40	3.51	1.33	16
APSyn*	1.73	3.69	1.74	16	0.55	1.41	0.55	4	1.38	3.02	1.39	10
balAPInc*	1.73	3.87	1.67	8	0.47	1.53	0.44	4	1.36	3.18	1.30	6
SLQS*	0.70	1.68	0.73	4	0.37	0.92	0.33	4	0.60	1.46	0.61	4
UMDuluth_C	8.13	18.93	7.53	20	-	-	-	-	-	-	-	-
EXPR_C	4.94	11.64	4.52	16	-	-	-	-	-	-	-	-
UMDuluth_E	-	-	-	-	3.79	9.99	3.66	28	-	-	-	-

Table 4: Results for the English subtask (1A). Baselines are marked with *, and those system participating only on Concepts or Entities are shown at the bottom and marked with either 'C' or 'E'.

tems can be used to speed up the hypernym discovery process, but they can also provide new hypernyms not considered beforehand.

Unsupervised distributional methods (e.g. the unsupervised baselines) seemed to perform poorly overall, as these systems tended to retrieve similar words which are not necessarily hypernyms. For example, false positives for APSyn and balAPInc are characterized by a large number of cohyponyms (e.g. *Exodus* and *Genesis*) and syntagmatically related words (e.g. *orange* and *juice*).

As regards the top performing systems, it is worth noting that they often tended to retrieve correct or near-correct hypernyms. The hypernyms that were retrieved on the gold standard were of several kinds: first, some hypernyms were present in the gold standard but normalized differently (for example, for *About.com* the gold standard contained *website* but not *web site* retrieved by CRIM_r1); second, they retrieved hy-

1B: Italian												
	Concepts				Entities				All			
	MAP	MRR	P@5	FPs	MAP	MRR	P@5	FPs	MAP	MRR	P@5	FPs
300-sparsans_r1	**8.94**	**18.77**	8.71	12	**22.56**	**46.34**	**21.79**	16	**12.08**	**25.14**	**11.73**	14
NLP_HZ	9.28	15.23	**9.12**	12	18.32	32.37	18.26	**28**	11.37	19.19	11.23	**20**
300-sparsans_r2	7.32	16.02	7.31	**16**	16.18	36.12	16.02	12	9.36	19.94	9.32	14
MFH*	5.07	13.30	4.31	0	16.71	39.56	15.18	8	7.76	19.37	6.82	4
vTE*	4.85	11.09	4.62	12	13.74	33.08	12.63	16	6.91	16.17	6.47	14
balAPInc*	4.84	10.71	4.84	**16**	0.72	1.96	0.77	4	3.89	8.69	3.90	10
APSyn*	4.30	9.50	4.33	12	1.00	2.06	1.00	4	3.54	7.56	3.56	8
SLQS*	2.02	4.02	2.07	4	0.26	0.75	0.17	0	1.62	3.26	1.63	2
Team 13	0.62	1.69	0.57	8	0.13	0.27	0.17	8	0.51	1.36	0.48	8

Table 5: Results for the Italian subtask (1B). Baselines are marked with *.

1C: Spanish												
	Concepts				Entities				All			
	MAP	MRR	P@5	FPs	MAP	MRR	P@5	FPs	MAP	MRR	P@5	FPs
NLP_HZ	**18.17**	25.17	**18.71**	12	23.19	33.48	23.21	**24**	**20.04**	28.27	**20.39**	**18**
300-sparsans_r1	13.21	**28.07**	12.80	8	**25.91**	**53.51**	**24.24**	4	17.94	**37.56**	17.06	6
300-sparsans_r2	11.10	22.90	11.07	**20**	14.92	30.87	15.14	12	12.52	25.87	12.59	16
MFH*	8.33	17.19	8.51	0	18.58	50.89	15.88	8	12.16	29.76	11.26	4
vTE*	6.08	14.32	6.01	12	8.84	20.96	9.10	4	7.11	16.80	7.16	8
balAPInc*	3.52	7.99	3.62	0	0.59	1.39	0.55	0	2.43	5.53	2.48	0
APSyn*	3.28	6.76	3.29	8	0.74	1.71	0.79	0	2.33	4.88	2.35	4
Team 13	2.57	6.08	2.06	12	0.06	0.13	0.05	4	1.63	4.31	1.65	8
SLQS*	1.21	2.27	1.14	0	0.37	0.89	0.32	0	0.90	1.75	0.83	0

Table 6: Results for the Spanish subtask (1C). Baselines are marked with *.

2B: Music												
	Concepts				Entities				All			
	MAP	MRR	P@5	FPs	MAP	MRR	P@5	FPs	MAP	MRR	P@5	FPs
CRIM_r1	**43.38**	**63.79**	**43.87**	24	38.42	55.54	38.76	12	**40.97**	**60.93**	**41.31**	16
CRIM_r2	41.98	63.07	42.32	20	34.59	51.08	35.80	8	40.88	60.18	41.58	16
MFH*	33.56	56.82	35.22	0	32.72	38.03	37.11	0	33.32	51.48	35.76	0
300-sparsans_r1	23.52	39.26	22.66	16	**44.71**	**64.53**	**44.48**	**20**	29.54	46.43	28.86	18
CRIM_CE	24.62	42.92	25.46	8	11.93	24.03	12.24	16	21.20	37.55	21.70	12
300-sparsans_r2	12.49	27.33	12.79	20	35.72	60.35	38.63	4	19.08	36.71	20.13	12
vTE*	11.53	35.78	10.28	12	16.67	48.39	17.77	**20**	12.99	39.36	12.41	16
Anu	10.68	27.13	10.84	**32**	3.43	7.19	3.90	8	8.62	21.47	8.87	**20**
vTE*_CE	6.31	16.54	6.81	4	13.37	33.58	14.87	4	3.51	9.79	3.62	4
SJTU BCMI	5.16	9.84	5.41	4	6.30	11.57	6.67	4	4.71	9.15	4.91	4
Team 13	4.83	14.33	4.51	12	2.82	7.92	3	8	5.62	16.87	5.11	16
ADAPT	1.88	5.34	1.89	2	0.00	0.00	0.00	0	2.63	7.46	2.64	4
balAPInc*	1.44	3.65	1.58	4	0.15	0.23	0.14	0	1.95	5.01	2.15	2
APSyn*	1.13	2.55	1.30	8	0.15	0.23	0.18	4	1.51	3.47	1.74	6
SLQS*	0.64	1.25	0.65	0	0.11	0.14		0	0.86	1.69	0.85	0

Table 7: Results for the Music subtask (2B). Baselines are marked with *.

pernyms which were either more or less fine-grained than the gold standard hypernyms (e.g. the list of gold hypernyms for *downfall* includes *natural phenomenon* but not *storm*, discovered by some supervised systems); third, some systems were able to retrieve hypernyms which correspond to another hyponym's sense not captured in the gold standard (e.g. *facultad* in Spanish can be either an educational institution or a virtue/ability, the latter not being captured by the gold standard

2A: Medical				
	MAP	MRR	P@5	FPs
CRIM_r1	**34.05**	**54.64**	**36.77**	20
CRIM_r2	31.54	46.19	35.49	12
MFH*	28.93	35.80	34.20	4
CRIM_CE	27.18	49.51	29.10	12
300-sparsans_r1	20.75	40.60	21.43	16
vTE*	18.84	41.07	20.71	12
300-sparsans_r2	14.96	32.18	15.81	12
EXPR_C	13.77	40.76	12.76	**40**
SJTU BCMI	11.69	25.95	11.69	12
vTE*_CE	11.66	23.83	12.64	32
ADAPT	8.13	20.56	8.32	20
Anu	7.05	17.51	7.29	32
Team 13	2.55	7.19	2.52	8
EXPR_C_CE	1.36	3.70	1.42	12
balAPInc*	0.91	2.10	1.08	0
APSyn*	0.65	1.43	0.72	4
SLQS*	0.29	0.66	0.33	0

Table 8: Results for the Medical subtask (2A). Baselines are marked with * and *cross evaluation* systems are followed by '_CE'.

but retrieved by the 300-sparsans_r2 system). Perhaps surprisingly, this latter case also extends to baselines such as MFH: in fact, many named entities have very skewed sense distributions, with less popular senses corresponding to people, cities, or companies often unbeknownst to most human annotators.[17] In addition to these three common patterns, there are also other correct false positives which do not clearly correspond to any of these three.

7 Conclusion

In this paper we have presented the SemEval 2018 task on *Hypernym Discovery*. We provided a large, reliable framework to evaluate hypernym discovery system in various languages (English, Italian, and Spanish) and domains (medical and music). This evaluation framework aims at going beyond the common practice of seeing hypernymy detection as a binary classification task, and provides a more challenging setting, inherently closer to how the task should be modeled within downstream applications. We hope this framework will contribute to the development of hypernym discovery systems in several languages and, more generally, to a wider understanding of hypernymy from a computational perspective.

As far as the results are concerned, this newly-proposed task proved to be challenging for all participating systems, leaving considerable room for improvement. It is clear from the figures that supervised systems perform considerably better than unsupervised systems. This might suggests that, given a well-defined downstream task, it could be more valuable to annotate hypernyms manually or semi-automatically (whenever possible) and then train a supervised system, than proposing unsupervised solutions with suboptimal performances. On the other hand, it is also noteworthy that the best system across three of the subtasks (i.e. CRIM) combined a supervised neural network architecture with the output of an unsupervised system using Hearst-style patterns (Hearst, 1992).

[17] As an example, *Cervantes* is universally known as the famous Spanish writer who authored '*Don Quixote*', but the word might also refer to a town in Western Australia.

Acknowledgements

The authors gratefully acknowledge the economic support in the construction of the datasets from the Maria de Maeztu-UPF Grant provided to Horacio Saggion, Luis Espinosa-Anke and Sergio Oramas; Google Research through the Google Doctoral Fellowship in Natural Language Processing to Jose Camacho-Collados; and Bar-Ilan University through Vered Shwartz. This work is partially supported by the TUNER project (TIN2015-65308-C5-5-R, MINECO/FEDER, UE), Spanish Ministry of Economy and Competitiveness.

We would also like to thank the task participants who provided helpful inputs to improve the task through their comments in the Google Group.

References

Marco Baroni, Raffaella Bernardi, Ngoc-Quynh Do, and Chung-chieh Shan. 2012. Entailment above the word level in distributional semantics. In *Proceedings of EACL*, pages 23–32.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.

Marco Baroni and Alessandro Lenci. 2011. How we blessed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 1–10. Association for Computational Linguistics.

Gbor Berend, Mrton Makrai, and Pter Fldik. 2018. 300-sparsans at semeval-2018 task 9: Hypernymy

as interaction of sparse attributes. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 925–931, New Orleans, Louisiana. Association for Computational Linguistics.

Gabriel Bernier-Colborne and Caroline Barriere. 2018. Crim at semeval-2018 task 9: A hybrid approach to hypernym discovery. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 722–728, New Orleans, Louisiana. Association for Computational Linguistics.

Guido Boella and Luigi Di Caro. 2013. Supervised learning of syntactic contexts for uncovering definitions and extracting hypernym relations in text databases. In *Machine learning and knowledge discovery in databases*, pages 64–79. Springer.

Gemma Boleda, Abhijeet Gupta, and Sebastian Padó. 2017. Instances and concepts in distributional space. In *Proceedings of EACL (2)*, Valencia, Spain. Association for Computational Linguistics.

Georgeta Bordea, Paul Buitelaar, Stefano Faralli, and Roberto Navigli. 2015. Semeval-2015 task 17: Taxonomy extraction evaluation (texeval). In *Proceedings of the SemEval workshop*.

Georgeta Bordea, Els Lefever, and Paul Buitelaar. 2016. Semeval-2016 task 13: Taxonomy extraction evaluation (texeval-2). In *SemEval-2016*, pages 1081–1091. Association for Computational Linguistics.

Jose Camacho-Collados. 2017. Why we have switched from building full-fledged taxonomies to simply detecting hypernymy relations. *arXiv preprint arXiv:1703.04178*.

Jose Camacho-Collados and Roberto Navigli. 2017. BabelDomains: Large-Scale Domain Labeling of Lexical Resources. In *Proceedings of EACL (2)*, Valencia, Spain.

Cristian Cardellino. 2016. Spanish Billion Words Corpus and Embeddings.

Luis Espinosa-Anke, Jose Camacho-Collados, Claudio Delli Bovi, and Horacio Saggion. 2016. Supervised distributional hypernym discovery via domain adaptation. In *Proceedings of EMNLP*, pages 424–435.

Tiziano Flati, Daniele Vannella, Tommaso Pasini, and Roberto Navigli. 2016. MultiWiBi: the Multilingual Wikipedia Bitaxonomy Project. *Artificial Intelligence*.

Ruiji Fu, Jiang Guo, Bing Qin, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Learning semantic hierarchies via word embeddings. In *Proceedings of ACL*.

Amit Gupta, Francesco Piccinno, Mikhail Kozhevnikov, Marius Pasca, and Daniele Pighin. 2016. Revisiting taxonomy induction over wikipedia. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2300–2309, Osaka, Japan.

Lushan Han, Abhay L Kashyap, Tim Finin, James Mayfield, and Jonathan Weese. 2013. Umbc_ebiquity-core: semantic textual similarity systems. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, volume 1, pages 44–52.

Arshia Zernab Hassan, Manikya Swathi Vallabhajosyula, and Ted Pedersen. 2018. Umduluth-cs8761 at semeval-2018 task 9: Hypernym discovery using hearst patterns, co-occurrence frequencies and word embeddings. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 911–915, New Orleans, Louisiana. Association for Computational Linguistics.

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

Johannes Hoffart, Dragan Milchevski, and Gerhard Weikum. 2014. Stics: searching with strings, things, and cats. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 1247–1248. ACM.

Ahmad Issa Alaa Aldine, Mounira Harzallah, Giuseppe Berio, Nicolas Bchet, and Ahmad Faour. 2018. Expr at semeval-2018 task 9: A combined approach for hypernym discovery. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 916–920, New Orleans, Louisiana. Association for Computational Linguistics.

Lili Kotlerman, Ido Dagan, Idan Szpektor, and Maayan Zhitomirsky-Geffet. 2010. Directional distributional similarity for lexical inference. *Natural Language Engineering*, 16(04):359–389.

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

Alfredo Maldonado and Filip Klubika. 2018. Adapt at semeval-2018 task 9: Skip-gram word embeddings for unsupervised hypernym discovery in specialised corpora. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 921–924, New Orleans, Louisiana. Association for Computational Linguistics.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

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 *ACL*, pages 1318–1327.

Kim Anh Nguyen, Maximilian Köper, Sabine Schulte im Walde, and Ngoc Thang Vu. 2017. Hierarchical embeddings for hypernymy detection and directionality. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 233–243, Copenhagen, Denmark.

Mihaela Onofrei, Ionut Hulub, Diana Trandabat, and Daniela Gifu. 2018. Apollo at semeval-2018 task 9: Detecting hypernymy relations using syntactic dependencies. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 895–899, New Orleans, Louisiana. Association for Computational Linguistics.

Sergio Oramas, Luis Espinosa-Anke, Mohamed Sordo, Horacio Saggion, and Xavier Serra. 2016. Elmd: An automatically generated entity linking gold standard dataset in the music domain. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association.

Sergio Oramas, Oriol Nieto, Francesco Barbieri, and Xavier Serra. 2017. Multi-label music genre classification from audio, text, and images using deep features. In *Proceedings of the 18th International Society of Music Information Retrieval Conference (ISMIR 2017)*.

Ellie Pavlick and Marius Pasca. 2017. Identifying 1950s american jazz musicians: Fine-grained isa extraction via modifier composition. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2099–2109. Association for Computational Linguistics.

Simone Paolo Ponzetto and Michael Strube. 2007. Deriving a large scale taxonomy from wikipedia. In *AAAI*, volume 7, pages 1440–1445.

John Prager, Jennifer Chu-Carroll, Eric W Brown, and Krzysztof Czuba. 2008. Question answering by predictive annotation. In *Advances in Open Domain Question Answering*, pages 307–347. Springer.

Wei Qiu, Mosha Chen, Linlin Li, and Luo Si. 2018. Nlp_hz at semeval-2018 task 9: a nearest neighbor approach. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 906–910, New Orleans, Louisiana. Association for Computational Linguistics.

Sara Rodríguez-Fernández, Luis Espinosa Anke, Roberto Carlini, and Leo Wanner. 2016. Semantics-driven recognition of collocations using word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 499–505.

Stephen Roller and Katrin Erk. 2016. Relations such as hypernymy: Identifying and exploiting hearst patterns in distributional vectors for lexical entailment. In *Proceedings of EMNLP*, pages 2163–2172, Austin, Texas.

Stephen Roller, Katrin Erk, and Gemma Boleda. 2014. Inclusive yet selective: Supervised distributional hypernymy detection. In *Proceedings of COLING 2014*, Dublin, Ireland.

V. Ivan Sanchez Carmona and Sebastian Riedel. 2017. How Well Can We Predict Hypernyms from Word Embeddings? A Dataset-Centric Analysis. In *Proceedings of EACL (short)*, pages 401–407.

Enrico Santus, Alessandro Lenci, Tin-Shing Chiu, Qin Lu, and Chu-Ren Huang. 2016a. Nine features in a random forest to learn taxonomical semantic relations. *arXiv preprint arXiv:1603.08702*.

Enrico Santus, Alessandro Lenci, Tin-Shing Chiu, Qin Lu, and Chu-Ren Huang. 2016b. Unsupervised measure of word similarity: How to outperform co-occurrence and vector cosine in vsms. In *Thirtieth AAAI Conference on Artificial Intelligence*, pages 4260–4261.

Enrico Santus, Alessandro Lenci, Qin Lu, and Sabine Schulte im Walde. 2014. Chasing hypernyms in vector spaces with entropy. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 38–42. Association for Computational Linguistics.

Vered Shwartz, Yoav Goldberg, and Ido Dagan. 2016. Improving hypernymy detection with an integrated path-based and distributional method. In *Proceedings of ACL*, Berlin, Germany.

Vered Shwartz, Enrico Santus, and Dominik Schlechtweg. 2017. Hypernyms under siege: Linguistically-motivated artillery for hypernymy detection. In *Proceedings of EACL*, Valencia, Spain. Association for Computational Linguistics.

Rion Snow, Daniel Jurafsky, and Andrew Y Ng. 2004. Learning syntactic patterns for automatic hypernym discovery. *Advances in Neural Information Processing Systems 17*.

Kent A Spackman, Keith E Campbell, and Roger A Côté. 1997. Snomed rt: a reference terminology for health care. In *Proceedings of the AMIA annual fall symposium*, page 640. American Medical Informatics Association.

Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: A core of semantic knowledge. In *WWW*, pages 697–706. ACM.

Aaron Swartz. 2002. Musicbrainz: A semantic web service. *IEEE Intelligent Systems*, 17(1):76–77.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Ivan Vulić, Daniela Gerz, Douwe Kiela, Felix Hill, and Anna Korhonen. 2017. Hyperlex: A large-scale evaluation of graded lexical entailment. *Computational Linguistics*.

Yogarshi Vyas and Marine Carpuat. 2017. Detecting asymmetric semantic relations in context: A case-study on hypernymy detection. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 33–43. Association for Computational Linguistics.

Chengyu Wang, Xiaofeng He, and Aoying Zhou. 2017. A short survey on taxonomy learning from text corpora: Issues, resources and recent advances. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1190–1203.

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.

J.C. Wu, Y.C. Chang, T. Mitamura, and J.S. Chang. 2010. Automatic collocation suggestion in academic writing. In *Proceedings of the ACL Conference, Short paper track*, Uppsala.

Mohamed Yahya, Klaus Berberich, Shady Elbassuoni, and Gerhard Weikum. 2013. Robust question answering over the web of linked data. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 1107–1116. ACM.

Zheng Yu, Haixun Wang, Xuemin Lin, and Min Wang. 2015. Learning term embeddings for hypernymy identification. In *Proceedings of IJCAI*, pages 1390–1397.

Zhousheng Zhang, Jiangtong Li, Hai Zhao, and Bingjie Tang. 2018. Sjtu-nlp at semeval-2018 task 9: Neural hypernym discovery with term embeddings. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 900–905, New Orleans, Louisiana. Association for Computational Linguistics.