# A Short Survey on Taxonomy Learning from Text Corpora: Issues, Resources and Recent Advances

**Chengyu Wang, Xiaofeng He*, Aoying Zhou**
Shanghai Key Laboratory of Trustworthy Computing,
School of Computer Science and Software Engineering, East China Normal University
chywang2013@gmail.com, {xfhe,ayzhou}@sei.ecnu.edu.cn

## Abstract

A taxonomy is a semantic hierarchy, consisting of concepts linked by is-a relations. While a large number of taxonomies have been constructed from human-compiled resources (e.g., Wikipedia), learning taxonomies from text corpora has received a growing interest and is essential for long-tailed and domain-specific knowledge acquisition. In this paper, we overview recent advances on taxonomy construction from free texts, reorganizing relevant subtasks into a complete framework. We also overview resources for evaluation and discuss challenges for future research.

## 1 Introduction

A taxonomy is a semantic hierarchy that organizes concepts by is-a relations, which exhibits the capability of improving many NLP and IR tasks, such as query understanding (Hua et al., 2017), personalized recommendation (Zhang et al., 2014), question answering (Yang et al., 2017), etc. It also supports a variety of real-world applications, including information management (Nickerson et al., 2013), e-commerce (Aanen et al., 2015) and biomedical systems (Köhler et al., 2014).

With massive Web data available, a number of taxonomies are constructed from human-compiled resources such as Wikipedia, Wikidata, etc (Suchanek et al., 2007; Ponzetto and Navigli, 2009; Flati et al., 2014; Mahdisoltani et al., 2015). But even large taxonomies may lack domain-specific and long-tailed knowledge. Recently, several methods have been developed to induce taxonomies from text corpora (Wu et al., 2012; Yang, 2012; Luu et al., 2014). However, this task is far from being solved for three reasons: i) Text corpora may vary in size, topic and quality. It is unlikely to develop a "one-size-fits-all" solution for all scenarios. For example, given an extremely large corpus, Hearst-pattern based method is employed to build Probase (Wu et al., 2012). For domain-specific corpora, learning hypernymy embedding is more suitable (Luu et al., 2016b). ii) The accuracy of free-text taxonomies is usually lower than many Wikipedia-based taxonomies because it is difficult to extract knowledge completely from texts; iii) The task of taxonomy learning is still insufficiently studied a) in emerging and specific domains and b) for non-English or under-resourced languages (Wei et al., 2014; Alfarone and Davis, 2015; Wang et al., 2015).

In this paper, we overview recent advances on taxonomy construction from text corpora, reorganizing relevant subtasks into a complete framework. The subtasks include hyponym acquisition, hypernym prediction, taxonomy induction, etc. We also summarize resources, evaluation metrics and state-of-the-art results. We also discuss issues and directions for future research.

## 2 Taxonomy Construction Techniques

Although workflows of different methods vary, a free text-based taxonomy construction system typically operates in two steps: i) extracting is-a relations using *pattern-based* (Sect. 2.1) or *distributional* methods (Sect. 2.2); ii) constructing a complete taxonomy from is-a relations (Sect. 2.3).

### 2.1 Pattern-based Methods

Traditional pattern based methods predict that there is an is-a relation between a term pair $(x, y)$, if $x$ and $y$ appear in the same sentence and satisfy a particular pattern. The earliest and most influential work in this field is Hearst (1992), which handcrafts several lexical patterns to harvest is-a relations. A typical pattern is "[C] such as [E]", where

*Corresponding author.



*Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1201–1214
Copenhagen, Denmark, September 7–11, 2017. 

[C] and [E] are placeholders of noun phrases that are regarded as the hypernym (class) $y$ and the hyponym (entity) $x$ respectively for an is-a relation $(x, y)$. Based on Hearst patterns, Probase is constructed from billions of Web pages. It consists of 2.65 million concepts and 20.76 million is-a relations (Wu et al., 2012). Similar approaches are presented in Etzioni et al. (2004); Kozareva and Hovy (2010), which employ Hearst patterns to induce taxonomies from Web pages.

Despite the successful applications, these patterns are too specific to cover all linguistic circumstances, thus recall is sacrificed (Wu et al., 2012). Simple pattern matching is prone to error due to idiomatic expressions, parsing errors, incomplete/uninformative extractions and ambiguous issues (Kozareva et al., 2008; Etzioni et al., 2004). In the next part, we summarize key techniques to improve precision and recall for pattern-based methods. Note that a robust is-a relation extraction system may combine multiple techniques to achieve high precision and recall.

#### 2.1.1 Methods Improving Recall

**Pattern Generalization** Several approaches either extend original Hearst patterns by linguistic rules or learn more generalized lexico-syntactic patterns. Ritter et al. (2009) increase recall by replacing the noun phrase "[E]" (i.e., candidate hyponym) in Hearst patterns with a list of $k$ noun phrases. Luu et al. (2014) design more flexible patterns where a few words in such patterns are interchangeable. Automatic methods mine is-a patterns given a collection of seed instances as input. Snow et al. (2004) use the dependency path of two terms to represent the pattern, where both syntactic and lexical connections of two terms can be modeled. This practice is more resistent to noise than surface matching and is employed by a number of relation extraction systems (Snow et al., 2006; Banko et al., 2007; Shwartz et al., 2016).

The number of patterns generated from a text corpus is sufficiently large, causing the feature sparsity problem. Learning more abstract patterns from these "raw" patterns can improve the generality of these patterns, hence increases recall. Navigli and Velardi (2010) introduce the concept "star pattern" (which use wildcards to replace non-frequent words in sentences). More general patterns are created by clustering star patterns. In the PATTY system (Nakashole et al., 2012), a subset of words along the dependency path are replaced by their POS tags, ontological types or wildcards.

**Iterative Extraction** Incorrect relations are frequently extracted from overly generalized patterns due to language ambiguity and *semantic drift* (Carlson et al., 2010). In contrast to above-mentioned approaches, an opposite idea is to use extremely specific patterns. Kozareva et al. (2008) employ "doubly-anchored" patterns (e.g., "cars such as Ford and *") to harvest hyponyms for a particular hypernym and expand both hyponyms and hypernyms by a bootstrapping loop. It uses each pattern as a query and takes search engine results as a Web corpus. Another advantage is that the ambiguity of terms can be eliminated by "doubly-anchored" patterns. Similar to Kozareva and Hovy (2010); Carlson et al. (2010), new is-a relations and hypernym patterns are iteratively extracted in an automatic manner.

**Hypernym Inference** This type of methods overcome the limitation where $x$ and $y$ must appear in the same sentence. The idea of Ritter et al. (2009) is that if $y$ is a hypernym of $x$ and another term $x^{'}$ is sufficiently similar to $x$, there is a high probability that $y$ is a hypernym of $x^{'}$. They train an HMM to learn a better similarity measure than vector-based approaches. In the Syntactic Contextual Subsumption (SCS) method (Luu et al., 2014), given a non-taxonomic relation $r$, denote $S_r(x)$ as the collection of objects such that for each $s \in S_r(x)$, $x$ and $s$ has the relation $r$. If $S_r(y)$ mostly contains $S_r(x)$ but not vice versa, we can infer $y$ is a hypernym of $x$.

Syntactic inference on hyponym modifiers can generate additional is-a relations. For example, the machine can infer a grizzly bear is a bear based on the evidence that the head word of "grizzly bear" is "bear". In Taxify (Alfarone and Davis, 2015), the system adds the linguistic head of a multi-word term as its direct hypernym if the term is added to the taxonomy. Suchanek et al. (2007) link conceptual Wikipedia categories to WordNet synsets based on category head words. Gupta et al. (2016) introduce linguistic heuristics to derive is-a relations from Wikipedia category network. Besides English, a similar observation also holds for Chinese, as shown in Fu et al. (2013); Li et al. (2015).

#### 2.1.2 Methods Improving Precision

**Confidence Assessment** After candidate is-a pairs $(x, y)$ are extracted, statistical measures can be used to estimate confidence scores. Relations

with low scores are discarded. In KnowItAll (Etzioni et al., 2004), the system estimates the pointwise mutual information (PMI) of $x$ and $y$ by search engine hit counts. Probase (Wu et al., 2012) employs the ratio of likelihood to determine the most possible hypernym $y$ for a concept $x$, and reversely the most possible hyponym $x$ for a concept $y$. Wu et al. (2012) further calculate the plausibility of extracted is-a pairs based on a Naive Bayes classifier. Besides statistics from extraction results, Luu et al. (2014, 2015) consider external factors, such as the inclusion of concepts in WordNet and dictionaries, as well as the trustworthiness of data sources (e.g. Web pages). The experience of building Google's Knowledge Vault (Dong et al., 2014) shows that assessing confidence scores is essential for acquiring and fusing knowledge from different extractors.

It is worth nothing that the negative evidence can be also employed to estimate confidence scores. A recent approach (Wang and He, 2016) uses statistics of both hypernym and co-hyponym patterns to give each pair a positive score and a negative score. Experiments show that using negative scores improves precision by discarding co-hyponym relations that are incorrectly predicted as is-a relations by their model.

**Classification-based Validation** These methods train a classifier $f$ to predict the correctness of an extracted pair $(x, y)$. Models of choice typically include SVM, logistic regression and neural nets. The features for $f$ can be roughly divided into following categories: surface name, syntax, statistics, external resources, etc. In the literature, Snow et al. (2004, 2006) use the dependency paths between $x$ and $y$ as features in the corresponding lexico-syntactic patterns. Ritter et al. (2009) introduce a list of features based the frequency of matches between a pair and Hearst patterns, such as the number of matches for "$x$ is a $y$" in a corpus. Surface name features (Bansal et al., 2014) consider the word formation of $x$ and $y$, including whether $x$ and $y$ are capitalized, whether $x$ ends with $y$. Bansal et al. (2014) further employ statistics derived from matches of Hearst patterns in the corpus and Wikipedia abstracts. This is because abstracts in Wikipedia contain definitions and summaries about concepts that can be used for inferring is-a relations.

Using both pattern-based and distributional representations of $x$ and $y$ can also enhance the performance of the classifier, as shown in Shwartz et al. (2016). This technique can be viewed as a combination of pattern-based and distributional methods, which will be discussed in details in Sect. 2.2.4.

## 2.2 Distributional Methods

Distributional methods predict is-a relations between terms based on their distributional representations, by either *unsupervised measures* (Sect. 2.2.2) or *supervised models* (Sect. 2.2.3). Because they directly *predict* is-a relations instead of *extracting* all is-a relations in a corpus, we briefly introduce how to obtain key terms to form term pairs as candidate is-a relations (Sect. 2.2.1).

### 2.2.1 Key Term Extraction

The first step for predicting is-a relations is to generate candidate hyponyms/hypernyms. For free texts, they are usually *key terms*, which are nouns, noun phases and/or named entities that frequently appear in the corpus. The key terms can be identified by applying POS tagging or NER tools to the corpora and then using rule-based extractors (Yang, 2012; Zhu et al., 2013; Luu et al., 2014). Existing keyword or key phrase extractors can be also used to recognize these terms automatically (Navigli et al., 2011; Qureshi et al., 2012; da Silva Conrado et al., 2013; Liu et al., 2015).

For learning domain-specific taxonomies, an important post-processing step after extracting key terms is *domain filtering*. This filters out terms not in the domain of interest, improving the taxonomy precision. A term can be filtered by statistics-based cuts, which include TF, TF-IDF, domain relevance, domain consensus (Navigli and Velardi, 2004; de Knijff et al., 2013) and domain specificity scores (Alfarone and Davis, 2015). To ensure the extracted terms are important concepts in a particular domain, several methods only harvest terms from domain definitive sentences (Navigli et al., 2011; Velardi et al., 2013; Anke et al., 2016b). Specially, Navigli et al. (2011) propose to use domain weights to select sentences that define unambiguous terms pertained to the domain of interest.

### 2.2.2 Unsupervised Measures

We first survey various unsupervised measures for is-a relation identification. After that, feature representations are introduced for these measures.

**Distributional Similarity Measures** Early work of *distributional similarity measures* mostly

focuses on *symmetric* measures such as cosine, Jaccard, Jensen-Shannon divergence and the widely cited LIN measure (Lin, 1998):

$$\text{LIN}(x,y) = \frac{\sum_{f \in F_x \cap F_y} w_x(f) + w_y(f)}{\sum_{f \in F_x} w_x(f) + \sum_{f \in F_y} w_y(f)}$$

where $x$ and $y$ are candidate hyponyms and hypernyms respectively. $F_x$ and $F_y$ are features of $x$ and $y$. $w_x(f)$ is the weight of feature $f$ for word $x$. But they only learn semantic similarity of words.

*Asymmetric* measures model the *asymmetric* property of is-a relations, following the *Distributional Inclusion Hypothesis* (DIH) (Geffet and Dagan, 2005; Zhitomirsky-Geffet and Dagan, 2009). It assumes that a hyponym only appears in some of its hypernym's contexts, but a hypernym appears in all contexts of its hyponyms. For example, the concept "fruit" has a broader spectrum of contexts than its hyponyms, such as "apple", "banana" and "pear". As an example, Weeds et al. (2004) propose a simple measure *WeedsPrec* to compute the weighted inclusion of features of $y$ within features of $x$:

$$\text{WeedsPrec}(x,y) = \frac{\sum_{f \in F_x \cap F_y} w_y(f)}{\sum_{f \in F_y} w_y(f)}$$

Other asymmetric measures are introduced in a variety of research, e.g., WeedsRec (Weeds et al., 2004), BalAPInc (Kotlerman et al., 2010), ClarkeDE (Clarke, 2009), cosWeeds, invCL (Lenci and Benotto, 2012), WeightedCosine (Rei and Briscoe, 2014). Detailed summarization of distributional similarity measures can be found in an early survey on vector space semantic models (Turney and Pantel, 2010).

More recently, several studies suggest that DIH is not correct for all the cases (Santus et al., 2014; Rimell, 2014). For example, "American" is a hypernym of "Barack Obama" but the (politics-related) contexts of "Barack Obama" cannot be covered by those of "American". Most contexts of a hypernym are less informative and more general than those of its hyponyms. To solve this problem, Santus et al. (2014) propose an entropy-based measure SLQS for hypernym detection. Roller et al. (2014) introduce the *Selective Distributional Inclusion Hypothesis*, which means the original DIH is correct, but only for relevant dimensions.

**Features** For each term $x$, a collection of features $F_x$ are generated from the text corpus, where each feature $f \in F_x$ represents a contextual word with which $x$ co-occurs (Lin, 1998; Weeds et al., 2004). In some work, $f$ also specifies the syntactic relation between $x$ and $f$ (Lin, 1998). As stated in Padó and Lapata (2003), the usage of syntactic-based vector space model can better distinguish different lexical relations than the simple "Bag-of-Words" co-occurrence model. In addition, Schütze (1993) use the context word and the position relative to the target term as features. Baroni and Lenci (2010) propose a Distributional Memory framework to generate word-link-word features. Yamada et al. (2009) use raw verb-noun dependencies and cluster such dependencies to generate feature vectors

The value of each feature is determined by a weight function $w_x(f)$, which quantifies the statistical association between the feature $f$ and the corresponding word $x$. Choices of $w_x(f)$ include the (point-wise) Mutual Information (PPMI) (Weeds et al., 2004), Local Mutual Information (LMI) (Evert, 2005). Dimension reduction methods such as SVD can be employed to create dense vectors (Roller and Erk, 2016).

### 2.2.3 Supervised Models

With training sets available, *classification/ranking methods* train a model to predict hypernymy based on the representations of a term pair $(x, y)$. *Hypernym generation* approaches directly model how to "generate" hypernyms based on the representations of hyponyms in the embedding space.

**Classification** In classification methods, the most popular representations for $x$ and $y$ are word embeddings generated by pre-trained neural language models such as Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014) and ivLBL (Mnih and Kavukcuoglu, 2013). SensEmbed (Iacobacci et al., 2015) generates different embeddings for different senses of the same word.

The *concat* model combines term-pair vectors by $\vec{x} \oplus \vec{y}$ where $\vec{x}$ is the embedding vector of word $x$, then trains an off-the-shelf classifier such as SVM (Baroni et al., 2012). This model is regarded as a strong baseline in some papers (Kruszewski et al., 2015; Shwartz et al., 2016; Mirza and Tonelli, 2016). Recent work points out that it has a serious problem of *lexical memorization* (Roller et al., 2014; Levy et al., 2015; Weeds et al., 2014). It means that the classifier learns the semantics of terms rather than the relations between the terms.

Consequently, when the training sets and testing sets are significantly different, the model suffers from a poor performance.

To over this problem, the *diff* model uses vector offsets as features, represented as $\vec{y} - \vec{x}$ (Rimell, 2014; Weeds et al., 2014; Fu et al., 2014). The *asym* model is presented in Roller et al. (2014), using both vector difference and squared vector difference features. The *simDiff* model (Turney and Mohammad, 2015) employs the difference of two word-context matrices (i.e., domain matrix and function matrix) as features for relation classification. Additionally, other combinations of vectors are mentioned in the literature, such as vector sum $\vec{x} + \vec{y}$, and dot-product $\vec{x} \cdot \vec{y}$ (Shwartz et al., 2016). Roller and Erk (2016) exploit Hearst patterns in distributional vectors and introduce a PCA-like iterative procedure to learn *concat* classifiers. Kruszewski et al. (2015) learn mappings from distributional vectors to boolean-valued vectors, where the output vectors correspond to entailment between words.

In neural language models (Mikolov et al., 2013; Pennington et al., 2014), words that occur in similar contexts have similar embeddings. Yu et al. (2015) argue that this modeling technique is not strong enough to learn term embeddings for is-a relation prediction. For each word $x$, they learn two types of embeddings $\vec{x}_o$ and $\vec{x}_e$, representing the embeddings of $x$ when $x$ functions as a hyponym and a hypernym, respectively. The embeddings are generated by training a distance-margin based neural net. Luu et al. (2016b) further extend this approach by modeling the contexts between hypernyms and hyponyms in a dynamic weighting neural net. Li et al. (2016) design a joint model based on negative sampling to embed entities and categories jointly into the same semantic space. The high performance shows that using task-specific embeddings is more effective than general-purpose embeddings.

**Hypernym Generation** *Hypernym generation* approaches make prediction for a pair $(x, y)$ by whether the model can map $\vec{x}$ to a vector close to $\vec{y}$. Fu et al. (2014) is a pioneer work in this field, which employs uniform linear projection and piecewise linear projection to map the embeddings of a hyponym to its hypernym. After that, three approaches (Wang and He, 2016; Yamane et al., 2016; Tan et al., 2015) have been proposed to extend Fu et al. (2014). Wang and He (2016) update transition matrices and extract new is-a relations iteratively. They improve the performance of the piecewise projection model when training sets and test sets have little overlap in the semantic space. Yamane et al. (2016) learn the number of clusters and transition matrices jointly by dynamically clustering is-a pairs. Tan et al. (2015) replace the transition matrix with the embedding of "is-a". As shown in Yamane et al. (2016), these methods are comparable to state-of-the-art classification approaches in terms of F-measure. Additionally, by domain clustering, the approach (Fu et al., 2014) is modified to a transfer learning version that is sensitive to target data sources for domain adaptation (Anke et al., 2016a).

The negative sampling technique proves effective to enhance projection learning. This is because the representations of hypernymy relations are sometimes confused with synonymy, co-hyponymy and meronymy. In Ustalov et al. (2017), an additional regularization term is added to the model of Fu et al. (2014) to take the advantage of the semantics of not-is-a relations. Wang et al. (2017) explicitly learn the representations of not-is-a relations so that the true hypernymy relations are better distinguished. This method considers the representations of both is-a and not-is-a relations, hypernym-level similarity and linguistic rules in a transductive learning setting.

**Ranking** As an alternative approach, Fu et al. (2013) present a ranking model to select the most probable hypernym for an entity. Replacing a classification model with a ranking model is not a common practice for extracting is-a relations, due to its low recall. However, this method is specifically engineered for the Chinese language. As shown in (Fu et al., 2014; Wang et al., 2015; Li et al., 2015; Wang and He, 2016), learning Chinese is-a relations is fundamentally challenging due to flexible language expressions. Thus it is necessary to train a ranking model to extract Chinese is-a relations with high precision.

### 2.2.4 Discussion

In the literature, there are some disagreements on which methods are more effective for is-a relation prediction. For example, Shwartz et al. (2016) claim that distributional methods outperform pattern-based approaches, while Levy et al. (2015) hold the opinion that distributional methods do not even work. We overview major view-

points in the research community and analyze pros and cons for both types of methods.

Pattern-based methods extract is-a relations $(x, y)$ based on the lexico-syntactic paths connecting $x$ and $y$ in a corpus, which explicitly express the relation. The original Hearst patterns (Hearst, 1992) and more generalized patterns have been used in a large number of taxonomies (Wu et al., 2012; Navigli et al., 2011). A disadvantage is that using patterns as features may result in the sparsity of the feature space (Nakashole et al., 2012). Most methods require $x$ and $y$ to co-occur in the same sentence. Hence, the recall is limited. Besides, they are overly language-dependent and difficult to use if there are few Hearst-like patterns in other languages. For example, as shown in Fu et al. (2014); Wang and He (2016), they suffer from extremely low recall for the Chinese language.

In contrast, distributional approaches use word representations derived from contexts, independent of its hyponym or hypernym. The usage of word embeddings (Mikolov et al., 2013) allows machines to make predictions based on the entire corpus. However, distributional methods are less precise in detecting specific, strict is-a relations and tend to discover broad semantic similarity between terms (Shwartz et al., 2016; Yu et al., 2015). As Weeds et al. (2014) discover, some terms detected by distributional methods are co-hyponyms and meronyms, rather than hypernyms. Another drawback is that the representations are domain dependent and the models are heavily related to the training set (Roller et al., 2014). Yet a further criticism is pointed out by Levy et al. (2015). They find that supervised distributional methods actually learn whether $y$ is a "prototypical hypernym", instead of the relation between $x$ and $y$. This is because the features $\vec{x}$ and $\vec{y}$ are generated independently. They integrate the intra-pair similarity with the *diff* model by kernel functions but only achieves an incremental improvement.

Despite their own disadvantages, pattern-based and distributional methods have been considered complementary. The idea of integrating them has been proposed early (Mirkin et al., 2006; Kaji and Kitsuregawa, 2008) but have not drawn much attention over the years. Recently, the HyperNET system (Shwartz et al., 2016) represents a pair $(x, y)$ by both distributional and pattern-based features. Each pattern is represented by a dependency path, and embedded by an LSTM model (Hochreiter and Schmidhuber, 1997). Experiments show that the joint representation improves the performance notably, having F1-scores of 0.901 and 0.700 over two large datasets. In contrast, the best pattern-based method (i.e., an extension of Snow et al. (2004)) has the performance of 0.761 and 0.660. The best distributional approach based on the *concat* model has the performance of 0.746 and 0.637. An extension of the previous model named LexNET (Shwartz and Dagan, 2016) is proposed to recognize multiple relations.

### 2.3 Taxonomy Induction

In this part, we summarize techniques for creating taxonomies from is-a relations.

**Incremental Learning** Several methods construct an entire taxonomy from a "seed" taxonomy via incremental learning. Snow et al. (2006) enrich WordNet by maximizing the probability of an extended taxonomy based on the evidence of is-a and cousin relations harvested from texts. They focus on extracting new entities and attaching them to the semantic hierarchy of WordNet. Shen et al. (2012) observe that the extracted terms can either refer to existing entities in the taxonomy or new ones, and propose a graph-based method to link these terms to the taxonomy or insert new entities into the taxonomy. While these methods rely heavily on existing taxonomies, Kozareva and Hovy (2010) take a root concept as input only and iteratively extract is-a relations to expand the taxonomy. Alfarone and Davis (2015) further consider the problem that a "seed" taxonomy cannot be obtained in a specific domain. They construct the "seed" taxonomy by Hearst pattern matching and heuristic rules.

**Clustering** Taxonomy learning can be modeled as a clustering problem where similar terms clustered together may share the same hypernym. Hierarchical clustering is employed to cluster similar terms into a taxonomy (Hjelm and Buitelaar, 2008; de Knijff et al., 2013; Meijer et al., 2014). Song et al. (2015) improve the hierarchical clustering technique by scalable Bayesian Rose Trees. A similar idea is also introduced in Alfarone and Davis (2015) where the lowest common ancestor of a collection of terms clustered by K-Medoids is inferred as their common hypernym. The SCS method (Luu et al., 2014) (see Sect. 2.1.1) is also related to clustering because it groups simi-

lar terms by non-taxonomic relations, and infer its hypernyms to improve the taxonomy coverage.

**Graph-based Induction** Graph-based approaches are naturally suitable for this task because taxonomies are generally graphs. Kozareva and Hovy (2010) derive the path from the root to a target term by finding the longest path in a raw graph where edges represent noisy is-a relations. Anke et al. (2016b) calculate the path weights by multiplying the domain pertinence scores of its edges. Another frequently used algorithm is the *optimal branching algorithm* (Velardi et al., 2013; Luu et al., 2014). It first assigns edge weights based on graph connectivity (e.g., in-degree, betweenness, etc.), and finds an optimal branching based on *Chu-Liu/Edmonds's algorithm* (Karp, 1971). After noisy edge removal, a rooted tree is constructed with maximum weights. Bansal et al. (2014) employ a factor graph model to represent terms and is-a relations. The learning of a taxonomy is regarded as a structured learning problem for the model, solved by loopy belief propagation.

**Taxonomy Cleansing** The final step of taxonomy learning is *taxonomy cleansing*, which removes wrong is-a relations to improve the quality. A recent study on Probase (Wu et al., 2012) shows that incorrect is-a relations may exist in taxonomies in the form of cycles (Liang et al., 2017a). By eliminating cycles, 74K wrong is-a relations are detected. This cycle removal process is also applied in Deshpande et al. (2013); Fu et al. (2014); Li et al. (2015).

Another issue is entity ambiguity. As discussed in Liang et al. (2017b), the transitivity property does not necessarily hold in automatically constructed taxonomies. For example, the facts "(Albert Einstein, is-a, professor)" and "(professor, is-a, position)" do not mean that "(Albert Einstein, is-a, position)". The ambiguity issue has been addressed in a few systems (Anke et al., 2016b; Wu et al., 2012; Ponzetto and Navigli, 2009) by word sense disambiguation. However, it is not fully solved. While learning multiple senses of the word "bank" (a financial institution or riverside) is easy nowadays, it is more challenging to distinguish whether the word "professor" refers to a particular person or a job title in the taxonomy learning process. Based on Liang et al. (2017b), we can safely conclude that there is a long way towards learning a fully-disambiguated taxonomy.

| Contributor/Paper | #Positive | #Negative |
|---|---|---|
| Kotlerman et al. (2010) | 1,068 | 2,704 |
| Baroni and Lenci (2011) | 1,337 | 13,210 |
| Baroni et al. (2012) | 1,385 | 1,385 |
| Jurgens et al. (2012) | 1,154 | 1,154 |
| Levy et al. (2014) | 945 | 11,657 |
| Rei and Briscoe (2014) | 3,074 | - |
| Weeds et al. (2014) | 2,564 | 3,771 |
| Turney and Mohammad (2015) | 920 | 772 |
| Shwartz et al. (2016) (Lex) | 5,659 | 22,636 |
| Shwartz et al. (2016) (Rnd) | 14,135 | 56,544 |

Table 1: Test sets for is-a relation prediction.

## 3 Resources and Analysis

In this section, we summarize resources and metrics for taxonomy learning. Results and recommendations for future research are also discussed.

### 3.1 Resources

There have been a variety of resources for the research of is-a relation prediction. The first type is high-quality taxonomies. knowledge bases and semantic networks. The knowledge in these systems can be used for generating training sets for distant supervised model learning. Typical English resources include WordNet (Miller, 1995), YAGO (Suchanek et al., 2007), WiBi (Flati et al., 2014) and DefIE (Bovi et al., 2015). For languages other than English, refer to multilingual systems (e.g., YAGO3 (Mahdisoltani et al., 2015), BabelNet (Navigli and Ponzetto, 2012), MultiWiBi (Flati et al., 2016)). We need to point out that these systems are not necessarily all taxonomies but contain rich type hierarchical knowledge. We also summarize some recent (2010∼) test sets and statistics in Table 1[1].

Two shared tasks are designed specifically for taxonomy learning, i.e., TExEval (SemEval-2015 Task 17) (Bordea et al., 2015) and TExEval-2 (SemEval-2016 Task 13) (Bordea et al., 2016)[2]. In TExEval, the goal is to construct taxonomies in four target domains (i.e. chemicals, equipment, food and science), each with gold-standard provided. The setting is expanded to cover four European languages (i.e., English, Dutch, French and

[1]Statistics combine training, validation and test sets. Papers that use subsets of these datasets are not listed. Dataset (Jurgens et al., 2012) is not directly capable of evaluating the task and is processed by Turney and Mohammad (2015).

[2]Due to the relatively large number of submissions, we do not provide citations to every system submitted to TExEval tracks. Readers can refer to the two reports for details.

Italian) in TExEval-2. In this task, the participants are encouraged to use the Wikipedia corpus as input but there are no restrictions on the data sources. In previous studies, several domain-specific corpora have also been employed as inputs for taxonomies, including AI papers (Velardi et al., 2013), biomedical corpora (Alfarone and Davis, 2015), Web pages related to animals, plants and vehicles (Kozareva et al., 2008) and MH370 (Luu et al., 2016a), terrorism reports (Luu et al., 2014), disease reports and emails (Luu et al., 2016a) and Wikipedia corpora related to specific topics.

### 3.2 Evaluation Metrics

Evaluating hypernymy prediction algorithms is by no means easy. Given a collection of is-a and not-is-a term pairs as ground truth, standard relation classification metrics such as Precision (P), Recall (R) and F-score (F) can be employed to make the comparison (Shwartz et al., 2016; Yu et al., 2015).

However, evaluating the quality of an entire taxonomy is a non-trivial task due to i) the large size of a taxonomy, ii) the difficulty of obtaining gold standard and iii) the existence of multiple aspects that should be considered such as topology, correctness and coverage. If gold standard taxonomies are available, denote $S = (V_S, E_S)$ and $G = (V_G, E_G)$ as the extracted and gold standard taxonomies where $V_S$ and $V_G$ are node sets, $E_S$ and $E_G$ are edge sets. Evaluation metrics introduced in the two shared tasks (Bordea et al., 2015, 2016) are briefly summarized as follows:

- Node coverage: $|V_S \cap V_G|$, $|V_S \cap V_G|/|V_G|$;
- Edge coverage: $|E_S \cap E_G|$, $|E_S \cap E_G|/|E_G|$, $(|E_S| - |E_S \cap E_G|)/|E_G|$;
- Edge correctness: $P = |E_S \cap E_G|/|E_S|$, $R = |E_S \cap E_G|/|E_G|$, $F = 2(P \cdot R)/(P + R)$;
- Cumulative Fowlkes&Mallows (Cumulative F&M) measure (Velardi et al., 2013).

The second type of metrics compares taxonomies generated by different methods. Size ($|V_S|$ and $|E_S|$) and quality are two factors to be considered. Human assessment is required to estimate the accuracy by sampling and labeling edges. Additionally, topological statistics, including the numbers of simple directed cycles, connected components and intermediate nodes, can check whether the taxonomy is a Direct Acyclic Graph (DAG) and well-structured.

### 3.3 Result Analysis and Discussion

While we have discussed is-a relation prediction in Sect. 2.2.4, we mostly focus on the overall performance for taxonomy learning in this part.

We first analyze results of the two shared tasks (Bordea et al., 2015, 2016) as they report performance of a variety of methods. In both task, two pattern-based methods (i.e., INRIASAC (Grefenstette, 2015) in TExEval and TAXI (Panchenko et al., 2016) in TExEval-2) consistently outperform others. INRIASAC uses frequency-based co-occurrence statistics, and substring inclusion heuristics to extract a set of hypernyms for hyponyms. TAXI crawls a domain-specific corpora and employs lexico-syntactic patterns and substrings for domain is-a relation extraction. However, the potential of distributional methods is not fully exploited as only one system uses such techniques. Besides, different systems may use their own corpora in these tasks and hence the results do not directly reflect the "goodness" of these algorithms. In multilingual tasks, there is a large decrease in performance w.r.t. other languages in TAXI. The research (Fu et al., 2014; Wang and He, 2016) shares a similar experience when several effective algorithms for English do not really work for Chinese. This phenomenon calls for language-specific algorithms for non-English language sources.

For other works, although knowledge sources and domains may differ, we notice that they suffer from the *low recall* problem. For example, recall values of Bansal et al. (2014); Luu et al. (2014); Navigli et al. (2011); Kozareva and Hovy (2010) are lower than 50% in most cases and domains. While improving precision is relatively easy by imposing constraints, increasing recall is more challenging because we aim to identify all is-a relations, no matter whether the relations are expressed explicitly or implicitly, in one or multiple sentences (Shwartz et al., 2016). This problem becomes severe when less focused and dynamic domains are considered (Velardi et al., 2013).

### 3.4 Our Recommendations

Based on our analysis, we discuss our recommendations to improve the performance of taxonomy learning that have not been sufficiently addressed.

**Ensemble Representations and Deep Architectures** Shwartz et al. (2016) show that combining pattern-based and distributional methods can

improve the performance of is-a relation extraction. We suggest that the performance can be further improved by studying how the two types of approaches reinforce each other. Neural network models (Yu et al., 2015; Luu et al., 2016b) are effective to learn deeper representations of both features. In our opinion, It is also possible to solve the problem put forward by Levy et al. (2015) by adding the information of semantic relatedness between term pairs mined from patterns to distributed representations.

Another related topic is that despite several embedding learning methods mentioned above, there is only limited success of deep learning paradigms for the taxonomy induction. We believe this is mostly because it is difficult to design a single objective for neural networks to optimize for this task. Hence how to take advantage of the deep learning boom for taxonomy induction is worth researching in the future.

**Benchmarks and Evaluation** Benchmarks for taxonomy learning are crucial to quantify the performance of state-of-the-arts. Benchmarks should contain text corpora, gold standards and evaluation metrics. Bordea et al. (2015, 2016) have provided some gold standard taxonomies in several domains and languages but do not require all the systems to run over the same corpus. Other works use standard test sets and data sources, as we have summarized in Sect. 3.1.

Several issues in current benchmarks and methods have already been pointed out by previous works. Levy et al. (2015) show that supervised systems actually over-perform due to the lexical memorization problem. Shwartz et al. (2017) suggest that unsupervised approaches are more robust than supervised methods but supervised methods outperform unsupervised ones. Camacho-Collados (2017) discuss whether the hypernymy detection task is indeed an appropriate task for evaluating is-a relations in the context of taxonomy learning systems or their integration in downstream applications. We can see that more in-depth research should be done towards a complete, widely-accepted benchmark for evaluation.

**Unambiguous and Canonicalized Terms** For lexical taxonomies, a term may have multiple surface forms and senses. The ambiguity issue makes taxonomy-based applications prone to error (Liang et al., 2017b). It is desirable to construct taxonomies where each node represents an unambiguous term associated with its possible surface forms and their contexts. In this way, the taxonomy automatically supports entity linking and is beneficial for IR applications (e.g., Web search) (Shen et al., 2015; Hua et al., 2017).

**Incorporating Domain Knowledge** Domain knowledge is essential for term and relation extraction in domain-specific corpora but it is difficult to obtain from such limited corpora. By exploiting facts derived from domain knowledge bases, a domain taxonomy would be learned via distant supervision and have higher coverage than existing methods (Alfarone and Davis, 2015). Thus, it is an important task to construct a taxonomy based on a text corpus and a knowledge base of a specific domain.

**Non-English and Under-resourced Languages** The task addressed in this paper has not been extensively studied for under-resourced languages. Specifically, pattern-based methods, although effective for English, are language-dependent to a large extent. How to apply existing approaches to languages that are significantly different from English (e.g., Chinese, Arabic and Japanese) is worthy of research.

## 4 Conclusion

In this paper, we present a survey on taxonomy learning from text corpora. We overview pattern-based and distributional methods to learn hypernymy from texts, and discuss how to induce taxonomies from is-a relations. While there is significant success, this task is still far from being solved. By addressing the issues discussed in this paper, we suggest that high-quality taxonomies can be constructed in more domains and languages, having a greater influence in NLP and IR research.

## Acknowledgements

We would like to thank the anonymous reviewers for their insightful comments. This work is partially supported by the National Key Research and Development Program of China under Grant No. 2016YFB1000904. Chengyu Wang would also like to thank the ECNU Outstanding Doctoral Dissertation Cultivation Plan of Action under Grant No. YB2016040 for the support of his research.

## References


Steven S. Aanen, Damir Vandic, and Flavius Frasincar. 2015. Automated product taxonomy mapping in an e-commerce environment. *Expert Syst. Appl.* 42(3):1298–1313.

Daniele Alfarone and Jesse Davis. 2015. Unsupervised learning of an IS-A taxonomy from a limited domain-specific corpus. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*. pages 1434–1441.

Luis Espinosa Anke, José Camacho-Collados, Claudio Delli Bovi, and Horacio Saggion. 2016a. Supervised distributional hypernym discovery via domain adaptation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. pages 424–435.

Luis Espinosa Anke, Horacio Saggion, Francesco Ronzano, and Roberto Navigli. 2016b. Extasem! extending, taxonomizing and semantifying domain terminologies. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. pages 2594–2600.

Michele Banko, Michael J. Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. 2007. Open information extraction from the web. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*. pages 2670–2676.

Mohit Bansal, David Burkett, Gerard de Melo, and Dan Klein. 2014. Structured learning for taxonomy induction with belief propagation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. pages 1041–1051.

Marco Baroni, Raffaella Bernardi, Ngoc-Quynh Do, and Chung-chieh Shan. 2012. Entailment above the word level in distributional semantics. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. pages 23–32.

Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics* 36(4):673–721.

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*. page 110.

Georgeta Bordea, Paul Buitelaar, Stefano Faralli, and Roberto Navigli. 2015. Semeval-2015 task 17: Taxonomy extraction evaluation (texeval). In *Proceedings of the 9th International Workshop on Semantic Evaluation*. pages 902–910.

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

Claudio Delli Bovi, Luca Telesca, and Roberto Navigli. 2015. Large-scale information extraction from textual definitions through deep syntactic and semantic analysis. *TACL* 3:529–543.

José Camacho-Collados. 2017. Why we have switched from building full-fledged taxonomies to simply detecting hypernymy relations. *CoRR* abs/1703.04178.

Andrew Carlson, Justin Betteridge, Richard C. Wang, Estevam R. Hruschka Jr., and Tom M. Mitchell. 2010. Coupled semi-supervised learning for information extraction. In *Proceedings of the Third International Conference on Web Search and Web Data Mining*. pages 101–110.

Daoud Clarke. 2009. Context-theoretic semantics for natural language: an overview. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*. page 112119.

Merley da Silva Conrado, Thiago Alexandre Salgueiro Pardo, and Solange Oliveira Rezende. 2013. A machine learning approach to automatic term extraction using a rich feature set. In *Proceedings of the 2013 Conference of the North American Chapter of the Association of Computational Linguistics*. pages 16–23.

Jeroen de Knijff, Flavius Frasincar, and Frederik Hogenboom. 2013. Domain taxonomy learning from text: The subsumption method versus hierarchical clustering. *Data Knowl. Eng.* 83:54–69.

Omkar Deshpande, Digvijay S. Lamba, Michel Tourn, Sanjib Das, Sri Subramaniam, Anand Rajaraman, Venky Harinarayan, and AnHai Doan. 2013. Building, maintaining, and using knowledge bases: a report from the trenches. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. pages 1209–1220.

Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. 2014. Knowledge vault: a web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pages 601–610.

Oren Etzioni, Michael J. Cafarella, Doug Downey, Stanley Kok, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. 2004. Web-scale information extraction in knowitall: (preliminary results). In *Proceedings of the 13th international conference on World Wide Web*. pages 100–110.

Stefan Evert. 2005. *The statistics of word cooccurrences: word pairs and collocations*. Ph.D. thesis, University of Stuttgart.

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*. pages 945–955.

Tiziano Flati, Daniele Vannella, Tommaso Pasini, and Roberto Navigli. 2016. Multiwibi: The multilingual wikipedia bitaxonomy project. *Artif. Intell.* 241:66–102.

Ruiji Fu, Jiang Guo, Bing Qin, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Learning semantic hierarchies via word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. pages 1199–1209.

Ruiji Fu, Bing Qin, and Ting Liu. 2013. Exploiting multiple sources for open-domain hypernym discovery. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. pages 1224–1234.

Maayan Geffet and Ido Dagan. 2005. The distributional inclusion hypotheses and lexical entailment. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*. pages 107–114.

Gregory Grefenstette. 2015. INRIASAC: simple hypernym extraction methods. In *Proceedings of the 9th International Workshop on Semantic Evaluation*. pages 911–914.

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

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

Hans Hjelm and Paul Buitelaar. 2008. Multilingual evidence improves clustering-based taxonomy extraction. In *Proceedings of the 18th European Conference on Artificial Intelligence*. pages 288–292.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9(8):1735–1780.

Wen Hua, Zhongyuan Wang, Haixun Wang, Kai Zheng, and Xiaofang Zhou. 2017. Understand short texts by harvesting and analyzing semantic knowledge. *IEEE Trans. Knowl. Data Eng.* 29(3):499–512.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. Sensembed: Learning sense embeddings for word and relational similarity. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*. pages 95–105.

David Jurgens, Saif Mohammad, Peter D. Turney, and Keith J. Holyoak. 2012. Semeval-2012 task 2: Measuring degrees of relational similarity. In *Proceedings of the 6th International Workshop on Semantic Evaluation*. pages 356–364.

Nobuhiro Kaji and Masaru Kitsuregawa. 2008. Using hidden markov random fields to combine distributional and pattern-based word clustering. In *Proceedings of the 22nd International Conference on Computational Linguistics*. pages 401–408.

Richard M. Karp. 1971. A simple derivation of edmonds' algorithm for optimum branchings. *Networks* 1(3):265–272.

Sebastian Köhler, Sandra C. Doelken, Christopher J. Mungall, Sebastian Bauer, Helen V. Firth, Isabelle Bailleul-Forestier, Graeme C. M. Black, Danielle L. Brown, Michael Brudno, Jennifer Campbell, David R. FitzPatrick, Janan T. Eppig, Andrew P. Jackson, Kathleen Freson, Marta Gîrdea, Ingo Helbig, Jane A. Hurst, Johanna Jähn, Laird G. Jackson, Anne M. Kelly, David H. Ledbetter, Sahar Mansour, Christa L. Martin, Celia Moss, Andrew Mumford, Willem Ouwehand, Soo-Mi Park, Erin Rooney Riggs, Richard H. Scott, Sanjay Sisodiya, Steven Van Vooren, Ronald J. Wapner, Andrew O. M. Wilkie, Caroline F. Wright, Anneke T. Vulto-van Silfhout, Nicole de Leeuw, Bert B. A. de Vries, Nicole L. Washington, Cynthia L. Smith, Monte Westerfield, Paul N. Schofield, Barbara J. Ruef, Georgios V. Gkoutos, Melissa Haendel, Damian Smedley, Suzanna E. Lewis, and Peter N. Robinson. 2014. The human phenotype ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Research* 42(Database-Issue):966–974.

Lili Kotlerman, Ido Dagan, Idan Szpektor, and Maayan Zhitomirsky-Geffet. 2010. Directional distributional similarity for lexical inference. *Natural Language Engineering* 16(4):359–389.

Zornitsa Kozareva and Eduard H. 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*. pages 1110–1118.

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

Germán Kruszewski, Denis Paperno, and Marco Baroni. 2015. Deriving boolean structures from distributional vectors. *TACL* 3:375–388.

Alessandro Lenci and Giulia Benotto. 2012. Identifying hypernyms in distributional semantic spaces. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*. pages 75–79.

Omer Levy, Ido Dagan, and Jacob Goldberger. 2014. Focused entailment graphs for open IE propositions. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*. pages 87–97.

Omer Levy, Steffen Remus, Chris Biemann, and Ido Dagan. 2015. Do supervised distributional methods really learn lexical inference relations? In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pages 970–976.

Jinyang Li, Chengyu Wang, Xiaofeng He, Rong Zhang, and Ming Gao. 2015. User generated content oriented chinese taxonomy construction. In *Proceedings of the 17th Asia-Pacific Web Conference*. pages 623–634.

Yuezhang Li, Ronghuo Zheng, Tian Tian, Zhiting Hu, Rahul Iyer, and Katia P. Sycara. 2016. Joint embedding of hierarchical categories and entities for concept categorization and dataless classification. In *Proceedings of the 26th International Conference on Computational Linguistics*. pages 2678–2688.

Jiaqing Liang, Yanghua Xiao, Yi Zhang, Seung-won Hwang, and Haixun Wang. 2017a. Graph-based wrong isa relation detection in a large-scale lexical taxonomy. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. pages 1178–1184.

Jiaqing Liang, Yi Zhang, Yanghua Xiao, Haixun Wang, Wei Wang, and Pinpin Zhu. 2017b. On the transitivity of hypernym-hyponym relations in data-driven lexical taxonomies. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. pages 1185–1191.

Dekang Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*. pages 296–304.

Jialu Liu, Jingbo Shang, Chi Wang, Xiang Ren, and Jiawei Han. 2015. Mining quality phrases from massive text corpora. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. pages 1729–1744.

Anh Tuan Luu, Siu Cheung Hui, and See-Kiong Ng. 2016a. Utilizing temporal information for taxonomy construction. *TACL* 4:551–564.

Anh Tuan Luu, Jung-jae Kim, and See-Kiong Ng. 2014. Taxonomy construction using syntactic contextual evidence. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. pages 810–819.

Anh Tuan Luu, Jung-jae Kim, and See-Kiong Ng. 2015. Incorporating trustiness and collective synonym/contrastive evidence into taxonomy construction. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. pages 1013–1022.

Anh Tuan Luu, Yi Tay, Siu Cheung Hui, and See-Kiong Ng. 2016b. Learning term embeddings for taxonomic relation identification using dynamic weighting neural network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. pages 403–413.

Farzaneh Mahdisoltani, Joanna Biega, and Fabian M. Suchanek. 2015. YAGO3: A knowledge base from multilingual wikipedias. In *Proceedings of the Seventh Biennial Conference on Innovative Data Systems Research*.

Kevin Meijer, Flavius Frasincar, and Frederik Hogenboom. 2014. A semantic approach for extracting domain taxonomies from text. *Decision Support Systems* 62:78–93.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 27th Annual Conference on Neural Information Processing Systems*. pages 3111–3119.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM* 38(11):39–41.

Shachar Mirkin, Ido Dagan, and Maayan Geffet. 2006. Integrating pattern-based and distributional similarity methods for lexical entailment acquisition. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*.

Paramita Mirza and Sara Tonelli. 2016. On the contribution of word embeddings to temporal relation classification. In *Proceedings of the 26th International Conference on Computational Linguistics*. pages 2818–2828.

Andriy Mnih and Koray Kavukcuoglu. 2013. Learning word embeddings efficiently with noise-contrastive estimation. In *Proceedings of the 27th Annual Conference on Neural Information Processing Systems*. pages 2265–2273.

Ndapandula Nakashole, Gerhard Weikum, and Fabian M. Suchanek. 2012. PATTY: A taxonomy of relational patterns with semantic types. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. pages 1135–1145.

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

Roberto Navigli and Paola Velardi. 2004. Learning domain ontologies from document warehouses and dedicated web sites. *Computational Linguistics* 30(2):151–179.

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

Roberto Navigli, Paola Velardi, and Stefano Faralli. 2011. A graph-based algorithm for inducing lexical taxonomies from scratch. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*. pages 1872–1877.

Robert C. Nickerson, Upkar Varshney, and Jan Muntermann. 2013. A method for taxonomy development and its application in information systems. *EJIS* 22(3):336–359.

Sebastian Padó and Mirella Lapata. 2003. Constructing semantic space models from parsed corpora. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*. pages 128–135.

Alexander Panchenko, Stefano Faralli, Eugen Ruppert, Steffen Remus, Hubert Naets, Cédrick Fairon, Simone Paolo Ponzetto, and Chris Biemann. 2016. TAXI at semeval-2016 task 13: a taxonomy induction method based on lexico-syntactic patterns, substrings and focused crawling. In *Proceedings of the 10th International Workshop on Semantic Evaluation*. pages 1320–1327.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. pages 1532–1543.

Simone Paolo Ponzetto and Roberto Navigli. 2009. Large-scale taxonomy mapping for restructuring and integrating wikipedia. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*. pages 2083–2088.

Muhammad Atif Qureshi, Colm O'Riordan, and Gabriella Pasi. 2012. Short-text domain specific key terms/phrases extraction using an n-gram model with wikipedia. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*. pages 2515–2518.

Marek Rei and Ted Briscoe. 2014. Looking for hyponyms in vector space. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*. pages 68–77.

Laura Rimell. 2014. Distributional lexical entailment by topic coherence. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. pages 511–519.

Alan Ritter, Stephen Soderland, and Oren Etzioni. 2009. What is this, anyway: Automatic hypernym discovery. In *Learning by Reading and Learning to Read, Proceedings of the 2009 AAAI Spring Symposium*. pages 88–93.

Stephen Roller and Katrin Erk. 2016. Relations such as hypernymy: Identifying and exploiting hearst patterns in distributional vectors for lexical entailment. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. pages 2163–2172.

Stephen Roller, Katrin Erk, and Gemma Boleda. 2014. Inclusive yet selective: Supervised distributional hypernymy detection. In *Proceedings of the 25th International Conference on Computational Linguistics*. pages 1025–1036.

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*. pages 38–42.

Hinrich Schütze. 1993. Part-of-speech induction from scratch. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*. pages 251–258.

Wei Shen, Jianyong Wang, and Jiawei Han. 2015. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Trans. Knowl. Data Eng.* 27(2):443–460.

Wei Shen, Jianyong Wang, Ping Luo, and Min Wang. 2012. A graph-based approach for ontology population with named entities. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*. pages 345–354.

Vered Shwartz and Ido Dagan. 2016. The roles of path-based and distributional information in recognizing lexical semantic relations. *CoRR* abs/1608.05014.

Vered Shwartz, Yoav Goldberg, and Ido Dagan. 2016. Improving hypernymy detection with an integrated path-based and distributional method. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*.

Vered Shwartz, Enrico Santus, and Dominik Schlechtweg. 2017. Hypernyms under siege: Linguistically-motivated artillery for hypernymy detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. page 6575.

Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. 2004. Learning syntactic patterns for automatic hypernym discovery. In *Proceedings of the 17th Annual Conference on Neural Information Processing Systems*. pages 1297–1304.

Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. 2006. Semantic taxonomy induction from heterogenous evidence. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*.

Yangqiu Song, Shixia Liu, Xueqing Liu, and Haixun Wang. 2015. Automatic taxonomy construction from keywords via scalable bayesian rose trees. *IEEE Trans. Knowl. Data Eng.* 27(7):1861–1874.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web*. pages 697–706.

Liling Tan, Rohit Gupta, and Josef van Genabith. 2015. USAAR-WLV: hypernym generation with deep neural nets. In *Proceedings of the 9th International Workshop on Semantic Evaluation*. pages 932–937.

Peter D. Turney and Saif M. Mohammad. 2015. Experiments with three approaches to recognizing lexical entailment. *Natural Language Engineering* 21(3):437–476.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *J. Artif. Intell. Res. (JAIR)* 37:141–188.

Dmitry Ustalov, Nikolay Arefyev, Chris Biemann, and Alexander Panchenko. 2017. Negative sampling improves hypernymy extraction based on projection learning. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. page 543550.

Paola Velardi, Stefano Faralli, and Roberto Navigli. 2013. Ontolearn reloaded: A graph-based algorithm for taxonomy induction. *Computational Linguistics* 39(3):665–707.

Chengyu Wang, Ming Gao, Xiaofeng He, and Rong Zhang. 2015. Challenges in chinese knowledge graph construction. In *Proceedings of the 31st IEEE International Conference on Data Engineering Workshops*. pages 59–61.

Chengyu Wang and Xiaofeng He. 2016. Chinese hypernym-hyponym extraction from user generated categories. In *Proceedings of the 26th International Conference on Computational Linguistics*. pages 1350–1361.

Chengyu Wang, Junchi Yan, Aoying Zhou, and Xiaofeng He. 2017. Transductive non-linear learning for chinese hypernym prediction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*.

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

Julie Weeds, David J. Weir, and Diana McCarthy. 2004. Characterising measures of lexical distributional similarity. In *Proceedings of the 20th International Conference on Computational Linguistics*.

Bifan Wei, Jun Liu, Jian Ma, Qinghua Zheng, Wei Zhang, and Boqin Feng. 2014. Motif-based hyponym relation extraction from wikipedia hyperlinks. *IEEE Trans. Knowl. Data Eng.* 26(10):2507–2519.

Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Qili Zhu. 2012. Probase: a probabilistic taxonomy for text understanding. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. pages 481–492.

Ichiro Yamada, Kentaro Torisawa, Jun'ichi Kazama, Kow Kuroda, Masaki Murata, Stijn De Saeger, Francis Bond, and Asuka Sumida. 2009. Hypernym discovery based on distributional similarity and hierarchical structures. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. pages 929–937.

Josuke Yamane, Tomoya Takatani, Hitoshi Yamada, Makoto Miwa, and Yutaka Sasaki. 2016. Distributional hypernym generation by jointly learning clusters and projections. In *Proceedings of the 26th International Conference on Computational Linguistics*. pages 1871–1879.

Hui Yang. 2012. Constructing task-specific taxonomies for document collection browsing. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. pages 1278–1289.

Shuo Yang, Lei Zou, Zhongyuan Wang, Jun Yan, and Ji-Rong Wen. 2017. Efficiently answering technical questions - A knowledge graph approach. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. pages 3111–3118.

Zheng Yu, Haixun Wang, Xuemin Lin, and Min Wang. 2015. Learning term embeddings for hypernymy identification. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*. pages 1390–1397.

Yuchen Zhang, Amr Ahmed, Vanja Josifovski, and Alexander J. Smola. 2014. Taxonomy discovery for personalized recommendation. In *Proceedings of the Seventh ACM International Conference on Web Search and Data Mining*. pages 243–252.

Maayan Zhitomirsky-Geffet and Ido Dagan. 2009. Bootstrapping distributional feature vector quality. *Computational Linguistics* 35(3):435–461.

Xingwei Zhu, Zhaoyan Ming, Xiaoyan Zhu, and Tat-Seng Chua. 2013. Topic hierarchy construction for the organization of multi-source user generated contents. In *Proceedings of the 36th International ACM SIGIR conference on research and development in Information Retrieval*. pages 233–242.